E

Easterlin Hypothesis

Diane J. Macunovich and Richard A. Easterlin

Abstract

The economic and social fortunes of a birth cohort tend to vary as a function of that cohort's relative size, approximated by the crude birth rate surrounding the cohort's birth. Effects have been observed on young men's earnings and unemployment rates, college enrolment rates, marriage and divorce, fertility, crime, and suicide rates. These effects have been found to be asymmetrical about the peak of a baby boom, and the original hypothesis has been extended to suggest a wide range of effects on the economy as a whole, from GDP growth rate, through interest rates and stock market performance, to measures of productivity.

Keywords

Aggregate demand; Cohort size effects; Crowding; Demographic transition; Easterlin hypothesis; Female labour force participation; Fertility; Inflation; Interest rates; Life cycle models; Marriage and divorce; Saving rates; Relative income; Relative cohort size; Productivity; GDP growth; Unemployment rates; College enrolment rates; Material aspirations; Preferences; Crime rates; Suicide rates

JEL Classifications

J11

The Easterlin, or 'relative cohort size', hypothesis as originally formulated posits that, other things constant, the economic and social fortunes of a cohort (those born in a given year) tend to vary as a function of its relative size, approximated by the crude birth rate surrounding the cohort's birth (Easterlin 1987). This hypothesis has since been extended to suggest a wider range of effects on the economy as a whole (Macunovich 2002).

Although cohort size effects were originally expected to be symmetrical around the peak of the baby boom, which in the United States entered the labour market around 1980, it is now thought that they are tempered by aggregate demand effects and by feedback effects from adjustments made by young adults on the 'leading edge' of a baby boom. As a result, cohorts – and the economy generally – on the 'leading edge of a baby boom fare much better than those on the 'trailing edge', when all else is equal.

The ultimate effects of changing relative cohort size are hypothesized to fall into these three categories:

1. Direct or first-order effects of relative cohort size on male relative income (the earnings of young men relative to their aspirations); male unemployment and hours worked; men's and women's college wage premium (the extra earnings of a college graduate relative to

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

those of a secondary school graduate); and levels of income inequality generally.
2. Second-order effects operating through male relative income, especially the demographic adjustments people make in response to changing relative income, such as changes in women's labour force participation and their occupational choices; men's and women's college enrolment rates; marriage and divorce; fertility; crime, drug use, and suicide rates; out-of-wedlock childbearing and the incidence of female-headed families; and living arrangements.
3. Third-order effects on the economy of changing relative cohort size and the resulting demographic adjustments, such as changes in average wage growth; the overall demand for goods and services in the economy and hence the growth rate of the economy; inflation, interest rates, and savings rates; stock market performance; industrial structure; measures of gross domestic product (GDP); and productivity measures.

The three categories of effect are discussed first in this article, followed by a consideration of feedback effects and a discussion of empirical evidence.

First-Order Effects

The linkage between higher birth rates and adverse social and economic effects arises from 'crowding mechanisms' operating within three major social institutions, the family, school and the labour market. Within the family, a sustained upsurge in the birth rate is likely to entail an increase in the average number of siblings, higher average birth order, and a shorter average birth interval, and there is a substantial literature in psychology, sociology and economics linking child development negatively to one or more of these magnitudes (Ernst and Angst 1983; Heer 1985). The negative effects that have been investigated range over a wide variety of phenomena. With regard to mental health, for example, there is evidence that problem behaviours such as fighting, breaking rules, and delinquency are associated with increased family size. Adverse effects on morbidity and mortality of children have been found to be associated with increased family size and shorter birth spacing. A negative association between IQ and number of siblings has been found in a number of studies, and, with IQ controlled for, between educational attainment and family size. The principal mechanism underlying such developments is likely to be the dilution of parental time and energy per child and family economic resources per child, associated with increased family size.

The family mechanisms just discussed imply that, on average, a larger cohort is likely to perform less well in school. But even in the absence of any adverse effects within the family, a large cohort is likely to experience crowding in schools, which reduces average educational performance (Freeman 1976). At any given time the human and physical capital stock comprising the school system tends to be either fixed in amount or to expand at a fairly constant rate, so that a surge in entrants into the school system tends to be accompanied by a reduction in physical facilities and teachers per student. In the United States, school planning decisions are divided among numerous local governments and private institutions, and expansion has tended to occur in reaction to, rather than in anticipation of, a large cohort's entry. Moreover, even when expansion occurs it is usually not accompanied by maintenance of curriculum standards, partly because of the diminishing pool of qualified teachers available to supply the needs of educational expansion.

The experience of a large cohort both in the family and in school is likely, in turn, to leave the cohort less well prepared, on reaching adulthood, for success in the labour market. But even if there were no prior effects, the entry of a large proportion of young and relatively inexperienced workers into the labour market creates a new set of crowding phenomena, because the expansion of complementary factor inputs is unlikely to be commensurate with that of the youth labour force. Additions to physical capital stock tend to be dominated by considerations other than the relative supply of younger workers, and the growth in

older, experienced, workers is largely governed by prior demographic conditions. Growth in the relative supply of younger workers results, in consequence, in a deterioration of their relative wage rates, unemployment conditions and upward job mobility (Welch 1979). The adverse effects of labour market crowding tend to reinforce those of crowding within the school and family. For example, the deterioration in relative wage rates of the young translates into lower returns to education and consequent adverse impact on school drop-out rates and college enrolment (Freeman 1976). Also, problems encountered in finding a good job may reinforce feelings of inadequacy or frustration already stirred up by some prior experiences at home or in school, and lead to lower labour force participation among young men.

Second-Order Effects

The relative economic standing of successive generations at a given point in time may be altered systematically by fluctuations in relative cohort size. If parents' living levels play an important role in setting their children's material aspirations, as socialization theory leads one to believe, then an increase in the shortfall of children's wage rates relative to parents will cause the children to feel relatively deprived and under greater pressure to keep up. The importance of relative status influences of this type in affecting attitudes or behaviour has been widely recognized in social science theory (Duesenberry 1949).

Confronted with the prospect of a deterioration in its living level relative to that of its parents, a large young adult cohort may make a number of adaptations in an attempt to preserve its comparative standing. Foremost among these are changes in behaviour related to family formation and family life (Macunovich and Easterlin 1990; McNown and Rajbhandary 2003). To avoid the financial pressures associated with family responsibilities, marriage may be deferred. If marriage occurs, wives are more likely to work and to put off childbearing. If a wife bears children, she is more likely to couple labour force participation with childrearing, and to have a smaller number of children more widely spaced (Macunovich 2002; Jeon and Shields 2005).

The process of demographic adjustment to changing relative income can best be thought of in terms of *ex ante* and *ex post* income; that is, the disposable per capita income of individuals prior to and then following the adjustments. Analyses of baby boom cohorts in the United States have found that a cohort's male relative income – individual earning potential of baby boomers relative to that of their parents – was significantly lower than the individual earning potential of *pre-boom* cohorts relative to *their* parents. But after making the type of demographic adjustments indicated above, the boomers managed to bring their per capita disposable income on a par with that of their parents (Easterlin et al. 1990).

Other reactions to the psychological stresses induced by large cohort size may be viewed as socially dysfunctional. Feelings of inadequacy and frustration, for example, may lead to disproportionate consumption of alcohol and drugs, to mental depression, and, at the extreme, to a higher rate of suicide (Pampel 2001; Stockard and O'Brien 2002). Feelings of bitterness, disappointment and rage may induce a higher incidence of crime (O'Brien et al. 1999). Within marriage, the stresses of conflicting work and motherhood roles for women, and feelings of inadequacy as a breadwinner for men, are likely to result in a higher incidence of divorce (Macunovich 2002). In the political sphere, the disaffection felt by a large cohort because of its lack of success may make it more responsive to the appeals of those who are politically alienated (O'Brien and Gwartney-Gibbs 1989).

Third-Order Effects

The second-order effects described in the previous section will, through reduced marriage rates and increased divorce and female labour force participation rates, reduce the proportion of households with stay-at-home spouses, which increases the tendency to purchase market replacements for the goods and services traditionally produced by

women in the home. The result is a 'commoditization' of many goods and services that used to be produced in the home. They are now exchanged in the market – and thus counted in official measures of GDP and productivity – whereas previously they were part of the excluded 'non-market' economy.

This commoditization of goods and services causes measures of industrial structure to skew strongly toward services and retail, away from agriculture and manufacturing, creating low-wage service jobs. In addition, the influx of inexperienced young workers as members of a large birth cohort – both men and women – into the labour market exacerbates any decline in productivity growth by changing the composition of the workforce to one dominated by inexperienced and therefore lower-productivity workers. This decline in relative wages of younger workers resulting from their oversupply would lead employers to substitute cheaper labour for more expensive capital, thus lowering the young workers' productivity still further by providing those low-wage workers with less productivity-enhancing machinery and technology.

Although some analysts maintain that the potential age structure effect of the baby boomers on personal savings is not large enough to explain the full drop in US national savings rates since the 1980s, studies of this phenomenon to date have focused only on the behaviour of the baby boomers themselves. However, one might argue that the baby boomers have affected the propensity to save in age groups other than their own. For example, because boomers' earnings were depressed and they experienced an inflated housing market when they went to buy homes (both the effects of their own large cohort size), many parents of baby boomers drew on their own savings in order to help with down payments.

When the age structure of children is permitted to affect consumption and savings, a very strong age-related pattern of expenditures and saving can be identified. Children induce savings on the part of their parents between the ages of five and 16, possibly in anticipation of later educational expenses. When the relationships identified in this way are combined with the changing age distribution in the US population during the 20th century, they produce a savings rate that fluctuates by plus or minus 25 per cent around the mean, simply as a result of changing age structure (Macunovich 2002).

Similarly, a strong effect has been identified of changing age structure (measured simply as the proportion of young to old in the population) on real interest rates and inflation, because of differential patterns of savings and consumption with age (McMillan and Baesel 1990). A higher proportion of young adults in a population will produce lower aggregate savings levels – and hence higher interest rates. In this model, today's lower interest and inflation rates are the result of the ageing of the baby boomers, as they begin to acquire assets for their retirement years. The converse of this phenomenon – the potential 'meltdown' effect of a retiring baby boom on financial markets, asset values and interest rates – has been described as well (Schieber and Shoven 1994).

Some research has estimated a strong effect of age structure on housing prices in the United States, with the entry of the baby boom into the housing market causing the severe house price inflation of the 1970s and 1980s, and the entry of the baby bust causing house price deflation (Mankiw and Weil 1989). Although some have disputed the magnitude of the effect estimated there, most researchers have confirmed its existence. A later study, for example, found significant effects of detailed (single year) age structure in the adult population on all forms of consumption, including housing demand, and on money demand (Fair and Dominguez 1991).

These potential effects on aggregate demand, savings rates, interest rates and inflation suggest that there might have been a connection between changing age structure and macroeconomic fluctuations in the United States and elsewhere during the 20th century. When the population of young adults is expanding, the resultant growth in demand for durable goods creates confidence in investors, while an unexpected slowdown in the growth rate of young adults could cause cutbacks in production and investment in response to inventory buildups, with a snowball effect throughout the economy. There was a close

correspondence in the United States in the 20th century between 'turnaround points' of growth in the key age group of 15–24, and significant economic dislocations in 1908, 1929, 1938 and 1974. Similarly, there was a correlation between age structure and economic performance in industrialized nations in the 1930s, and in both industrialized and developing nations since the 1980s, with the 'Asian Tigers' some of the most recent examples (Macunovich 2002).

Feedback Effects on the Relationship Between Relative Cohort Size and Relative Income

Easterlin's original statements recognized the potential effects of outside influences on the relative cohort size mechanism (Easterlin 1987). However, the dynamic nature of the mechanism – the fact that many of these other factors would, in fact, be secondary and tertiary results of changing relative cohort size, and thus endogenous in any empirical application – has not been fully appreciated in most analyses to date. As a result, it is often concluded that the hypothesis may have been relevant in the post-Second World War period up to about 1980, but that it fails to extend beyond one full cycle to apply to the period since 1980.

The aggregate demand effect of changing relative cohort size, discussed in the previous section, is hypothesized to contribute significantly to the observed asymmetry in relative cohort size effects on male relative income. Although cohorts on the leading edge of a baby boom experience declining wages relative to those of older workers, they do so in an economy experiencing strong growth in aggregate demand resulting from the increasing relative cohort size among young adults. Cohorts on the lagging edge of a baby boom, however, enter a labour market weakened by the economic slump resulting from a transition from expanding to contracting relative cohort size.

Similarly, as one of the secondary effects of changing relative cohort size discussed earlier, female labour force participation is hypothesized to have increased in response to declining male relative income as the leading edge of the baby boom entered the labour market. If, as hypothesized, these young women also increased their levels of educational attainment in anticipation of future labour market participation, they would have in many cases competed directly with the male members of their cohort and exacerbated the effects of relative cohort size on male relative income. This effect would have been greatest for cohorts on the lagging edge of the boom – those who should have benefited from declining relative cohort size. It is important in empirical analyses to recognize the potential endogeneity of these other factors, rather than treat relative cohort size effects as 'contingent' on exogenous changes in female labour force participation, educational attainment and wages. Wage analyses based on relative cohort size which control for a cohort's position in the US baby boom – and thus allow for aggregate demand and female labour force changes – can explain most of the observed change in young men's entry level wages and in their returns to experience and education (Macunovich 2002).

Empirical Analyses

Empirically, the most important application of the hypothesis has been to explain the varying experience of young adults in the United States since the Second World War. There is, however, some evidence of its relevance to the experience of developed countries more generally in this period (Korenman and Neumark 2000; Pampel 2001; Stockard and O'Brien 2002; Jeon and Shields 2005), and perhaps as a mechanism leading to fertility decline during the demographic transition in developing countries (Macunovich 2002).

Overall, however, empirical analyses testing various aspects of the Easterlin hypothesis have produced fairly mixed results. By 2007 there have been two comprehensive analyses of the literature on the Easterlin hypothesis, and one meta-analysis of 19 studies completed between 1976 and 2002. The meta-analysis (Waldorf and Byun 2005) focused on the age structure–fertility link, and concluded that analytical problems contribute to

an apparent lack of empirical support for the Easterlin hypothesis. Most significant among these were the failure to recognize the endogeneity of an income variable when combined with a relative cohort size variable, and the use of very broad age groups in defining relative cohort size.

The first of the literature reviews considered a broad range of topics, including labour market experience and education; marriage, fertility and divorce; and crime, suicide and alienation. It concluded:

> [T]he evidence for the Easterlin effect proves mixed at best and plain wrong at worst... Aggregate data support the hypothesis more than individual level data, period-specific or time-series data support the hypothesis more than cohort-specific data, experiences from 1945–1980 support the hypothesis more than the years since 1980, and trends in the United States support the hypothesis more than trends in European nations. (Pampel and Peters 1995, p. 189.

The second literature review evaluated 76 published analyses focused solely on fertility, and concluded:

> With an equal number of micro- and macro-level analyses using North American data (twenty-two), the 'track record' of the hypothesis is the same in both venues, with fifteen providing significant support in each case. The literature suggests unequivocal support for the relativity of the income concept in fertility but is less clear regarding the source(s) of differences in material aspirations, and suggests that the observed relationship between fertility and cohort size has varied across countries and time periods due to the effects of additional factors not included in most models. (Macunovich 1998, p. 53)

This review suggests that, because of data limitations and idiosyncratic interpretations of the hypothesis by individual researchers, many of the studies with unfavourable findings have been only peripherally related to the Easterlin hypothesis.

Conclusion

Since the early 1980s, demographic concepts have encroached modestly on economic theory, as evidenced by the appearance of life cycle, overlapping generations and vintage models. The cohort size hypothesis might be viewed as another in this sequence. Its roots, however, extend beyond economics, reaching out into sociology, demography and psychology, and it seeks to encompass a wider range of attitudinal and behavioural phenomena than is traditionally considered economic.

See Also

- ▶ Demographic Transition
- ▶ Economic Demography

Bibliography

Duesenberry, J.S. 1949. *Income, saving, and the theory of consumer behaviour*. Cambridge, MA: Harvard University Press.

Easterlin, R.A. 1980. *Birth and fortune*, 1st ed. New York: Basic Books.

Easterlin, R.A. 1987. *Birth and fortune*, 2nd ed. Chicago: University of Chicago Press.

Easterlin, R.A., C. Macdonald, and D.J. Macunovich. 1990. How have the American baby boomers fared? Earnings and well-being of young adults 1964–1987. *Journal of Population Economics* 3: 277–290.

Ernst, C., and J. Angst. 1983. *Birth order: Its influence on personality*. Berlin: Springer.

Fair, R.C., and K. Dominguez. 1991. Effects of the changing U.S. age distribution on macroeconomic equations. *American Economic Review* 81: 1276–1294.

Freeman, R.B. 1976. *The overeducated American*. New York: Academic Press.

Heer, D.M. 1985. Effects of sibling number on child outcome. *Annual Review of Sociology* 11: 27–47.

Jeon, Y., and M.P. Shields. 2005. The Easterlin hypothesis in the recent experience of higher-income OECD countries: A panel-data approach. *Journal of Population Economics* 18: 1–13.

Korenman, S., and D. Neumark. 2000. Cohort crowding and the youth labour market: A cross-national analysis. In *Youth employment and joblessness in advanced countries*, NBER comparative labour market series, ed. D.G. Blanchflower and R.B. Freeman. Chicago: University of Chicago Press.

Macunovich, D.J. 1998. Fertility and the Easterlin hypothesis: An assessment of the literature. *Journal of Population Economics* 11: 1–59.

Macunovich, D.J. 2002. *Birth quake: The baby boom and its after shocks*. Chicago: University of Chicago Press.

Macunovich, D.J., and R.A. Easterlin. 1990. How parents have coped: The effect of life cycle decisions on the economic status of pre-school age children, 1964–1987. *Population and Development Review* 16: 301–325.

Mankiw, N.G., and N.D. Weil. 1989. The baby boom, the baby bust and the housing market. *Regional Science and Economics* 19: 235–258.
McMillan, H.M., and J.B. Baesel. 1990. The macroeconomic impact of the baby boom generation. *Journal of Macroeconomics* 12: 167–195.
McNown, R., and S. Rajbhandary. 2003. Time series analysis of fertility and female labour market behaviour. *Journal of Population Economics* 16: 501–523.
O'Brien, R.M., and P.A. Gwartney-Gibbs. 1989. Relative cohort size and political alienation: Three methodological issues and a replication supporting the Easterlin hypothesis. *American Sociological Review* 54: 476–480.
O'Brien, R.M., J. Stockard, and L. Isaacson. 1999. The enduring effects of cohort characteristics on age-specific homicide rates 1960–1995. *American Journal of Sociology* 104: 1061–1095.
Pampel, F.C. 2001. *The institutional context of population change: Patterns of fertility and mortality across high-income nations*. Chicago: University of Chicago Press.
Pampel, F.C., and H.E. Peters. 1995. The Easterlin effect. *Annual Review of Sociology* 21: 163.
Schieber, S.J., and J.B. Shoven. 1994. *The consequences of population aging on private pension fund saving and asset markets*, Working paper no. 4665. Cambridge, MA: NBER.
Stockard, J., and R.M. O'Brien. 2002. Cohort effects on suicide rates: International variations. *American Sociological Review* 67: 854–872.
Waldorf, B., and P. Byun. 2005. Meta-analysis of the impact of age structure on fertility. *Journal of Population Economics* 18: 14–40.
Welch, F. 1979. Effects of cohort size on earnings: The baby boom babies' financial bust. *Journal of Political Economy* 87: 65–97.

East-West Economic Relations

Marie Lavigne

The decade 1966–1975 is usually considered as the golden age of East-West economic relations. Already during the previous decade, i.e. since the end of the cold war, the USSR and the Eastern European countries had increased their trade with the West at an annual rate of growth slightly higher than their total trade. But after 1966 the expansion of trade and cooperation was sustained both by a favourable political climate and by strong economic complementarities between the West (here equated to the OECD countries) and the East (the USSR and the six European countries that are members of the CMEA, or Council for Mutual Economic Assistance; hereafter we shall mention them as CPEs or centrally planned economies, for the sake of brevity).

These years were marked by *détente*, initiated in 1966 with the triumphal visit to the USSR of the French President General de Gaulle. This was not only a bilateral event, but it set the stage for diversified and institutionalized links between Eastern and Western European economies. Later on, in 1972, US President Nixon's visit to Moscow opened the shorter phase of bright USUSSR economic relations which ended in 1975. At the beginning of that year, the Soviet Union unilaterally repudiated the Soviet-American treaty of commerce, as a retaliation for the deprivation of the most favoured nation clause; according to the American legislation just introduced, the clause could not be granted to a country restricting the rights for its citizens to emigrate. Before *détente* came altogether to its end, it was symbolically magnified in the final Act of the Conference for Security and Co-operation in Europe, signed in Helsinki in August 1975. The economic 'basket' of this text was meant to appear as the Charter of East-West mutually profitable relations.

From the economic point of view, the 1966–1975 decade was indeed a time of converging interests. The USSR and Eastern European countries had just engaged in economic reforms. They needed to modernize their industries. The Western firms found new markets for selling equipment and turnkey plants. High rates of economic growth, both in the West and in the East, sustained the prospects for increased exports from the East to the West, once the new capacities acquired from the West were put into operation. An era of deepening industrial cooperation, based upon technology imports and reverse flows of manufactured goods, seemed to open.

It was then almost forgotten that even in such a favourable context, East-West trade accounted for less then 3% of world trade. While in 1975 it amounted to slightly under 30% of total trade for the CPEs (slightly more for the USSR and less for the six smaller CPEs taken together), it never

exceeded 5% of total trade for the Western countries, except for some non-typical cases (such as Austria or Finland).

The following decade, ending in 1985, has witnessed a general shrinking of East-West trade. There was a conspicuous deterioration of the political climate with the invasion of Afghanistan by the Soviet troops in December 1979 and, 2 years later, martial law in Poland. The world economic crisis exerted some adverse effects as well. True, it benefited the USSR as an oil exporter. But the Western recession hampered the export drive of the smaller CPEs. The manufactured goods which they intended to export so as to repay their imports of equipment became less saleable in the East. Thus the imbalance between imports and exports, which had been steadily growing since 1970, could not be corrected through expanded sales. An easier way out was to borrow on Western financial markets. The CPEs were still creditworthy, and the level of international liquidity was high as a result of the inflow of petro-dollars. The total indebtedness of the CPEs culminated in 1981. The subsequent adjustments conducted in 1981–1983 (though a decrease in imports and domestic investment) ended up with a marked improvement in the CPEs external financial position and with a decrease in their foreign debt (except for Poland). But the general slowdown of growth in the East, partly due to these adjustments, does not allow for a steep upward trend in East-West trade.

The outlook for East-West economic relations is to be evaluated through the combination of two opposed sets of factors. On the one hand, there are strong interests on both sides pressing for the expansion of trade and cooperation. On the other, equally strong obstacles are hindering such a development. The outcome is probably to be seen in a stabilization of those relations, below the level reached during the 'golden age' decade.

Economic Interests

East-West trade is sometimes said to be a one-way street. As the magnitudes of shares in total trade show, these relations are several times more important for the East than for the West. However, dependencies are to be found on both sides, with an uneven distribution.

In the West, European countries are the main group of partners. They account for roughly 75% of sales to the East and 90% of imports from the East (figures of 1983). This pattern has been stable since the end of the 1970s. In 1970 the share of Western Europe was very similar on the import side, but larger on the export side (about 10 points more). Since then, two major exporters have emerged outside Europe, Japan (for technology) and the United States (for grain, mainly to the USSR).

In the East, the USSR gained a growing share of East-West trade after 1970. From twofifths of the total trade of the European CPEs with the West, it reached 50% in the mid-1970s and over 60% in the 1980s. This is mainly due to the increase in oil prices after 1973; it allowed the Soviet Union to secure a higher rate of growth of its trade with the West compared with the other CPEs up to 1980, and to avoid the decrease in trade which the other CPEs experienced in the beginning of the 1980s.

This growing concentration of East-West trade on the Soviet Union is an expression of stronger interdependences.

For Western Europe, especially for the large industrial corporations, the USSR emerged in the 1970s as a major purchaser of heavy equipment, whose orders helped to sustain the level of activities and jobs during the recession years. The controversial multi-billion dollars gas pipeline deal concluded in 1981 is a clear demonstration of such interests. When in 1982 the US government tried to oppose the supply of tubes and other equipment for the pipeline, as a retaliation for the Soviet role in the Polish crisis, and also as an attempt to reduce the export capacities of natural gas of the USSR, the European governments backed their firms. Even though the Soviet orders for equipment substantially declined after then, the Soviet Union remains a huge market.

On the other side, the Soviet Union has become a significant supplier of energy to Western Europe. Fuels now account for about 80% of its sales to the West, from about half that share in the

beginning of the 1970s. The major Western European energy importers (Germany, France, Italy) are now dependent for 6–7% of their total energy imports on the Soviet Union. For natural gas alone, their dependence may be above the 30% mark at the end of the 1980s, from about 15% to 20% a decade earlier. The Soviet market is a means of achieving a diversification in energy imports; it is a cheaper supplier for oil and gas because of the distance factor, and may be considered as a more reliable one, than the Third World.

Regarding trade with the United States, the major link is grain. The Soviet Union began to buy large quantities of American grain in 1975–1976 and has been the largest single customer of the United States since then. US sales never again reached the 70% share of Soviet grain imports which they formed in 1979. However, the strength of economic versus political interests is clearly demonstrated by the failure of the grain embargo, which had to be lifted under the pressure of American farmers. The long-term grain sales agreement linking the two countries, first signed in 1975, has not only been renewed but also supplemented with an anti-embargo clause (in 1983).

The Western trade of Eastern Europe lacks these powerful interdependences. The smaller CPEs taken together are on average less involved in trade with the West than the USSR. In 1984, the share of Western trade in their total trade was about 25% (against 30% for the USSR) and had been declining since 1980. But while Bulgaria and Czechoslovakia, much more oriented toward trade within CMEA, have a very low share of their total trade with the West (12–15%), Hungary (35%), Poland, GDR and Romania (30%) are potentially interested in expanding their trade with the West. However, opportunities for that are low. Their supply is made of sensitive goods (steel, chemicals, textiles, manufactured goods, agricultural products), the demand for which is sluggish in the West – and they complain of growing protectionism. For these goods competition is growing on Western markets from the new industrializing countries of the Third World, which in addition are more advanced in some high technology fields (electronics). They can hardly expect concessions from Western countries, for which they provide less promising markets than the USSR. The development of compensation deals is only a marginal way of securing outlets for their goods.

Obstacles

In the background of these differentiated economic interests, specific obstacles hinder East-West trade, in the political, institutional (systemic) and financial fields, to which must be added the 1986 developments on the world oil market.

Is East-West trade *political* in essence? In Western Europe, politics and economic relations are regarded as distinct by governments and firms. The lasting failure to find an agreement between EEC and the CMEA, since the beginning of official talks in 1976, is mainly due to the lack of institutional competence of the CMEA in matters of trade as appraised by the EC Commission (even if on the side of the Commission there is a political concern to avoid strengthening the Soviet-dominated CMEA as an organization). The major involvement of politics in East-West economic relations is related to US policy. The 'linkage' concept of tying economic advantages to Soviet concessions in the political sphere was associated in the 1970s with commercial policies (the granting of the MFN clause) or financial conditions (for access to bank credits). Since the end of that decade it has evolved into a policy of sanctions, first as a retaliation for the Soviet invasion of Afghanistan in 1979 (the grain embargo against the USSR, which was lifted in April 1981, and a tighter control of high technology sales); then as a response to the martial law imposed in December 1981 in Poland. In this last case the sanctions hit Poland (though credit and export restrictions, a suspension of the MFN clause), and the USSR (through attempts to stop the Eurosiberian pipeline deal by preventing the Western European countries from selling equipment to the USSR and from concluding the agreements for the purchases of gas). They were also

extended to the other CPEs through a very severe credit squeeze. All these measures culminated in 1982. They proved largely ineffective but generated conflicts within the Western Alliance. The major and lasting field of political pressure is to be found in the embargo on high technology sales to the CPEs, conducted through the Cocom (Coordinating Committee), an informal organization set up in 1949 and including the NATO countries plus Japan. Very active during the years of the cold war, it seemed to be withering in the late 1970s but regained momentum from 1980 on. The present rationale of the Cocom restrictions is threefold: to impose sanctions; to prevent the Soviet bloc from acquiring dual-use technologies (for military as well as civilian ends); to enlarge the scope of controls by restricting high-technology exports of non-Cocom members (Sweden, Switzerland, Austria, and even some Third World countries such as India).

The *systemic* obstacles in trade are related to the specific organization of state trading in the CPEs. The monopoly of foreign trade and the related planning of trade flows remain very rigid in the Soviet Union. Increased flexibility has been introduced in the trade mechanisms of all the other CPEs, where enterprises are gaining easier access to foreign trade transactions. Direct interfirm contacts have been stimulated through industrial cooperation. In all these countries except for GDR, it is now possible to create joint enterprises with foreign equity capital (the experiences remain limited). The state trading system, however reformed, still prevents the CPEs from successfully adjusting to the market requirements in the West.

The *financial* problems of East-West relations are less dramatic than in 1980–1981, when the total indebtedness of the USSR and Eastern Europe combined exceeded $80 billion, more than four times its level of the end of 1974. Two countries, Poland and Romania, entered in 1981 a process of rescheduling, which is still going on for Poland. Two others, GDR and Hungary, successfully managed to restore their external accounts in 1982–1984. Since then, the Western banks have again been ready to expand their loans not only to the Soviet Union, which has always remained a good risk, but also to the other CPEs, which by all accounts seem more creditworthy than the Third World.

East-West economic relations are finally to be replaced in the broader context of the CPEs' foreign economic relations, including intra-CMEA trade. The move toward closer integration, advocated by the Soviet Union at the Summit meeting of the CMEA in June 1984 and based upon the heavy requirements of the USSR regarding its imports from its partners, might well appear as an additional constraint to the expansion of East-West relations for the smaller CPEs.

The fall in oil prices, since the end of 1985, may have strong adverse effects on East-West trade. If the average price for oil is for some time stabilized at half its 1985 level, the Soviet Union will lose at least one third of its export gains in its trade with the West. These losses may be compensated for, in the short run, by cuts in imports and increased borrowing, together with a stronger pressure on the smaller CMEA countries. The latter will thus have to divert to the Soviet market goods exportable to the West. In addition, they too will lose as sellers of refined oil products, with the same consequences as for the USSR. The 'golden age' of East-West trade is definitely not to be renewed.

See Also

- ▶ Convergence Hypothesis
- ▶ Cycles in Socialist Economies
- ▶ Socialist Economies

References

Bornstein, M., Z. Gitelman, and W. Zimmerman (eds.). 1981. *East-west relations and the future of eastern Europe: Politics and economics*. London: Allen and Unwin.

Economic Bulletin for Europe. 1949 onwards. Geneva: Economic Commission for Europe, United Nations. Each volume contains developments on East-West trade.

Fallenbuchl, Z., and C. McMillan (eds.). 1980. *Partners in east-west relations, the determinants of choice*. New York: Pergamon Press.

Holzman, F. 1976. *International trade under communism, politics and economics*. New York: Basic Books.
Lavigne, M. 1979. *Les relations économiques Est-Ouest*. Paris: Presses Universitaires de France.
Lavigne, M. 1985. *Economie internationale des pays socialistes*. Paris: Armand Colin.
Levcik, F. (ed.). 1978. *International economics – Comparisons and interdependencies. Essays in honour of F. Nemschak*. Vienna: Springer.
Marer, P., and J.M. Montias (eds.). 1980. *East European integration and east-west trade*. Bloomington: Indiana University Press.

Eckstein, Otto (1927–1984)

Lester C. Thurow

Keywords
Eckstein, O.; Forecasting; Macroeconometric models

JEL Classifications
B31

Eckstein was an entrepreneur who moved a whole technology from the research community into the marketplace. Until he founded Data Resources, Inc., macroeconometric models were research vehicles and not vehicles for aiding business decision making. Under his direction Data Resources came to dominate the marketplace for this type of information, but more importantly it changed the nature of the game. To be taken seriously after his innovation, all economic forecasts had to be buttressed with econometric equations and no large firm would attempt to begin its decision-making processes without an understanding of the national and international economic forecasts emanating from such models.

Born in Ulm, Germany, in 1927, Dr. Eckstein fled to England in 1938 and came to the United States in 1939. He graduated from Stuyvesant High School in New York City and served in the United States Army Signal Corps from 1946 to 1947. He received an AB degree from Princeton University in 1951 and a Ph.D. from Harvard University in 1955.

In 1968, he and Donald B. Marron founded Data Resources, Inc., which has grown into the largest economic information company in the world. The firm became a subsidiary of McGraw-Hill, Inc. in 1979. He directed the development of the Data Resources Model of the US economy, and was responsible for its forecasting operations.

As an immigrant to the United States from Nazi Germany, Otto Eckstein wanted to contribute something to America's future success. Better economic policies that would lead to a higher American standard of living were not an abstraction to him. They were the centre of his professional life.

His professional career began with the analysis of large scale multi-year water resources projects and how one might better allocate national resources in such projects. In the late 1950s he was the principal intellectual director of a Joint Economic Committee study on how the United States might break out of what was then seen as the stagnation of the mid-1950s. His study on growth, full employment and price stability laid the basis for the successful economic policies that were followed in the first two-thirds of the 1960s. But he went on to implement those intellectual foundations as a member of the President's Council of Economic Advisers under President Johnson.

No one who knew the enthusiasm of Otto Eckstein for studying, teaching, and practising economics could thereafter think of economics as the dismal science.

Selected Works

1958a. Water resources development: *The economics of project evaluation. Cambridge, MA: Harvard University Press.*
1958b. (With J.V. Krutilla.) *Multiple purpose river development.* Baltimore: Johns Hopkins Press.
1964a. *Public finance.* New York: Prentice-Hall. 4th edn, 1979.

1964b. (With E.S. Kirschen and others.) *Economic policy in our time.* Amsterdam: North-Holland.
1967. (ed.) Studies in the economics of income maintenance. Washington, DC: Brookings.
1970. (ed.) *The econometrics of price determination.* Washington, DC: Board of Governors of the Federal Reserve System and Social Science Research Council.
1976. (ed.) Parameters and policies in the U.S. economy. Amsterdam: North- Holland.
1978. *The great recession.* Amsterdam: North-Holland.
1981. *Core inflation.* New York: Prentice-Hall.
1983. *The DRI model of the U.S. economy. New York: McGraw-Hill.*
1984. (With C. Caton, R. Brinner and P. Duprey.) The DRI report on U.S. manufacturing industries. New York: McGraw-Hill.

Ecole Nationale des Ponts et Chaussées

Robert B. Ekelund Jr and Robert F. Hébert

French School of Civil Engineering, located at 28 rue des Saint-Pères, Paris. Established in 1747 by Daniel Trudaine, Finance Minister to Louis XV, the Ecole has traditionally produced economists of exceptional talent and originality. Isnard, Dupuit and Cheysson were students there and at various times its faculty included the likes of Henri Navier, Joseph Minard, Joseph Garnier, Henri Baudrillart, Charles Gide, Clément Colson and François Divisia.

The idea of an institution dedicated to the professionalization of French engineers had its roots in the 17th century. In 1690 Vauban created the Corps of Military Engineers, which was to serve as a model for future public bodies of this sort. He even went so far as to propose a public examination to test the scientific knowledge of young people aspiring to become engineers. After an inauspicious beginning, the Ecole slowly acquired more scholarly aspirations. It was directed in its formative years by J.R. Perronet, who established the high standards and pedagogical technique responsible for the ultimate success of the school, so much so that French engineers became the envy of the world. Although a formal course in economics was not established until 1847 (receiving impetus from Dupuit's pioneer researches in 1844), engineers were 'officially' exhorted to study economics as early as 1792.

The Revolution of 1789 brought sweeping changes to higher education in France. For a time it seemed as though the Ecole would be swept away as a vestige of the *ancien régime,* but Mirabeau successfully defended its existence, and by the time Napoleon came to power, a major expansion of faculty, students and curriculum was under way. With the establishment of the Ecole Polytechnique in 1794, the nature of the Ecole des Ponts et Chaussées changed from an undergraduate to a postgraduate institution, offering admission by competitive examination and specialized training for polytechnicians seeking to become civil engineers. These civil engineers became problem-solvers in the areas of flood control, municipal water distribution and sewage disposal, canal building, railway construction, road building and myriad other matters of concern to engineers.

The 19th century was the 'golden age' of the Ecole, a time when the faculty was upgraded and the curriculum was stretched to include stereotomy (1799), modern languages (1806), mineralogy and geology (1826), administrative law (1831), political economy (1847), thermodynamics (1851), and applied chemistry (1864). The role of the Ecole was pivotal and international in both engineering and economics. Henri Navier, for example, was sent in 1821 and in 1823 by the Director General of the Corps to study British achievements in suspension bridge design and construction. Upon his return Navier, who wrote a number of essays on the economic value of public works, offered a *Mémoire sur les ponts suspendus* which brought the French to the forefront of such technology for much of the 19th century. Jules Dupuit entered the Ecole in 1824, where he reacted to both Navier's engineering and

economic studies, later producing a theory of marginal utility and a full scale welfare analysis of markets and market structure. In 1830 an American student, Charles Ellet, Jr., entered the Ecole, returning home as the premier suspension bridge builder and designer of his age *and* as one of the most creative American economic theorists of the century. In short, the 19th century is the period when economic inquiry at the Ecole burgeoned, easily outdistancing the policy squabbles that occupied French academic economists at the universities and in academic journals. It was the era of Dupuit, Cheysson and Colson, the unrecognized giants of 19th century French economics.

Today the Ecole des Ponts et Chaussées stands as the oldest of France's *grandes écoles*. Perched at the top of a rigid and highly centralized educational system, it persists in admitting the country's intellectual elite and in providing them with solid training in economics.

References

Ekelund Jr., R.B., and R.F. Hébert. 1973. Public economics at the Ecole des Ponts et Chaussées: 1830–1850. *Journal of Public Economics* 2(3): 241–256.

Ecological Economics

Anastasios Xepapadeas

Abstract

Ecological economics is the study of the interactions and co-evolution in time and space of human economies and the ecosystems in which human economies are embedded. It uncovers the links and feedbacks between human economies and ecosystems, and so provides a unified picture of ecology and economy. The link between ecology and human economies has been manifested in the development of resource management or bio-economic models, in which the main focus has been on fishery or forestry management where the impact of humans on ecosystems is realized through harvesting. More closed links have been developed, however, as both disciplines evolve.

Keywords

Adiabatic parameter; Biodiversity; Bioeconomics; Boulding, K.; Diffusion; Eecological economics; Ecology; Ecosystems; Externalities; Fisher–Kolmogorov equation; Intrinsic growth rate; Kolmogorov model; Logistic function; Lotka, A.; Nash equilibrium; Optimality behaviour; Predator–prey models; Random walk models; Red Queen hypothesis; Spatial economics; Tragedy of the commons; Turing mechanism; Volterra, V

JEL Classifications

B5; Q2

Ecology can be regarded as the study of living species such as animals, plants and microorganisms, and the relations among them and their natural environment. In this context, an ecosystem includes these species and their non-living environment, their interactions, and their evolution in time and space (see, for example, Roughgarden et al. 1989). Economics, meanwhile, is the study of how human societies use scarce resources to produce commodities and to distribute them among their members.

The need for an interdisciplinary approach – 'ecological economics' – stems from the fact that natural ecosystems and human economies are closely linked. In the process of production and consumption, human beings use ecosystems and their services, influence their evolution, and are the recipients of feedbacks originating from their actions upon ecosystems. As Kenneth Boulding (1965) notes in his classic paper 'Earth as a space ship', which can be regarded as a landmark in the emergence of ecological economics, 'Man is finally going to have to face the fact that he is a biological system living in an ecological system, and that his survival

power is going to depend on his developing symbiotic relationships of a closed-cycle character with all the other elements and populations of the world of ecological systems.'

Thus, ecological economics can be regarded as the study of the interactions and co-evolution in time and space of human economies and the ecosystems in which human economies are embedded. This implies that the task of ecological economics is to bridge the gap between economy and ecology by uncovering the links and the feedbacks between human economies and ecosystems, and by using these links and feedbacks to provide a unified picture of ecology and economy and their interactions and co-evolution. In a sense, ecological economics aims at linking ecological models and economic models in order to provide insights into complex and interrelated phenomena stemming from and affecting both ecosystems and human economies.

The natural link between ecology and human economies has been manifested in the traditional development of resource management or bio-economic models (for example, Clark 1990), in which the main focus has been on fishery or forestry management where the impact of humans on ecosystems is realized through harvesting. More close links have been developed, however, as both disciplines evolve.

Common methodological approaches may also be encountered in ecology and economics. Optimality behaviour, which is fundamental in economics, has also been used to provide insights into the structure of ecological systems, in the context of optimal foraging behaviour, species competition, or net energy maximization by organisms (for example, Tschirhart 2000; Tilman et al. 2005) with the purpose of founding macrobehaviours in ecosystems – such as those emerging from population dynamics – on microfoundations.

In the same context, the classical phenomenological-descriptive approach to species competition based on Lotka–Volterra systems has recently been complemented by mechanistic resource-based models of species competition for limiting resources (Tilman 1982, 1988). This approach has obvious links to competition among economic agents for limited resources. Furthermore, by linking the functioning of natural ecosystems with the provision of useful services to humans, or by using concepts such as ecosystems productivity, insurance from the genetic diversity of ecological systems against catastrophic events, or development of new products using genetic resources existing in natural ecosystems (Heal 2000), new insights into the fundamental issues of the valuation of ecosystems or the valuation of biodiversity have been derived. (Examples of useful services to humans include provisioning services, such as food, water, fuel, genetic material; regulation services, such as climate regulation, disease regulation; and cultural services and supporting services, such as soil formation, nutrient cycling; see Millennium Ecosystem Assessment 2005.)

Ecological Models

The traditional bio-economic models (Clark 1990), which describe the evolution of the population or the biomass of species when harvesting takes place, have formed the building blocks of ecological-economic modelling. These models can be extended along various lines to provide a more realistic picture of ecosystems (for a detailed analysis, see Murray 2003) and help build meaningful ecological-economic models. To start with, let $x(t)$ denote the biomass of a certain species at time t. Then evolution of the biomass is described by an ordinary differential equation

$$\frac{dx(t)}{dt} = \text{birth} - \text{naturaldeath} + \text{migration} - \text{harvesting}. \quad (1)$$

In the analysis of population models it is common, unless it is a specific case, to set the migration rate at zero, and to represent the natural rate of population growth (birth-natural death) by a function $F(x)$. The most common specification of this function is the famous *logistic function*, which is

$F(x) = rx(1-x/K)$. In this function r is a positive constant called *intrinsic growth rate* and K is the *carrying capacity* of the environment which depends on factors such as resource availability or environmental pollution. If we denote by $h(t)$ the rate of harvesting of the species biomass by humans, the population model becomes:

$$\frac{dx(t)}{dt} = F(x) - h(t), x(0) = x_0. \quad (2)$$

If $h(t) \equiv F(x)$, the population remains constant and the harvesting rate corresponds to *sustainable yield*. Harvesting rate is usually modelled as population dependent or $h = qEx$, where q is a positive constant, referred to as a *catchability coefficient* in fishery models, and E is *harvesting effort*. Human activities can affect the species population, in addition to harvesting, by affecting parameters such as the intrinsic growth rates or the carrying capacity. For example, if the stock of environmental pollution of a certain pollutant (such as phosphorus in a lake) in a natural ecosystem is denoted by P, with dynamics described by

$$\frac{dP(t)}{dt} = g(s(t), P(t)), P(0) = P_0, \quad (3)$$

where $s(t)$ is the rate of emissions (such as phosphorus loadings), and the pollutant affects parameters of the population model, then the combined model will be (3) along with

$$\frac{dx(t)}{dt} = r(P)x\left[1 - \frac{x}{K(P)}\right] - qEx, r'(P) < 0, K'(P) < 0. \quad (4)$$

If the catchability coefficient is affected by technical change, then it can be expressed by a function of time as $q(t)$. In this case (4) is not autonomous. Alternatively q can be a function of technological variables like R&D evolving in the economic module.

The population model (2) can be generalized to age-structured populations and multi-species populations. In multi-species populations the Lotka–Volterra predator–prey models are classic. If we denote the prey population by $x(t)$ and the predator population by $y(t)$ and ignore harvesting for the moment to simplify things, the model can be written as

$$\frac{dx(t)}{dt} = x\left[r\left(1 - \frac{x}{K}\right) - yR(x)\right], x(0) = x_0 \quad (5)$$

$$\frac{dy(t)}{dt} = ym\left(1 - \frac{ny}{x}\right), y(0) = y_0 m, n > 0 \quad (6)$$

where $R(x)$ is a function called the *predation term*, which can be specified as $\gamma x/(x^2+\delta^2)$, γ, $\delta > 0$. A more general multi-species model with J prey and J predators can be written, for $i = 1,\ldots,J$, as

$$\frac{dx_i(t)}{dt} = x_i\left[a_i - \sum_{j=1}^{J}\beta_{ij}y_i\right], x_i(0) = x_{i0} \quad (7)$$

$$\frac{dy_i(t)}{dt} = y_i\left[\sum_{j=1}^{J}\gamma_{ij}x_j - \delta_i\right], y_i(0) = y_{i0} \quad (8)$$

where all parameters are positive constants. An even more general model of interacting populations can be obtained by the generalized Kolmogorov model where the evolution of each species biomass is described by:

$$\frac{dx_i(t)}{dt} = x_i F_i(x_1, x_2, x_3, \ldots) i = 1, 2, 3, \ldots \quad (9)$$

In the mechanistic resource-based models of species competition emerging from the work of Tilman (for example, Tilman 1982, 1988), species compete for limiting resources. (For the use of this model in ecological-economic modelling, see Brock and Xepapadeas 2002; Tilman et al. 2005.) In these models the growth of a species depends on the limiting resource, and interactions among species take place through the species' effects on the limiting resource. Let $\mathbf{x} = (x_1,\ldots,x_n)$ be the vector of species biomasses and R the amount of the available limiting resource. Then a mechanistic resource-based model with a single limiting factor in a given

area and $i = 1,\ldots,n$ species can be described by the following equations:

$$\frac{\dot{x}_i}{x_i} = g_i(R) - d_i, x_i(0) = x_{i0} \tag{10}$$

$$\dot{R} = S - aR - \sum_{i=1}^{n} w_i x_i g_i(R) \tag{11}$$

where $g_i(R)$ is resource-related growth, d_i is the species' natural death rate, S is the amount of resource supplied, a is the natural resource removal rate (leaching rate), and w_i is specific resource consumption by species i. The main result in this framework relates to an exclusion principle stating that, in a landscape free of disturbances, the species with the lowest resource requirement in equilibrium will competitively displace all other species, driving the system to a monoculture. Species coexistence and polycultures in equilibrium can be supported in a system with more than one limiting resource, or even in single resource systems if there is temperaturedependent growth and temperature variation in the ecosystem, spatial or temporal variations in resource ratios, differences in local palatabilities and local abundance of herbivores.

In addition to the temporal variation captured by the models described above an important characteristic of ecosystems is that of spatial variation. Biological resources tend to disperse in space under forces promoting 'spreading' or 'concentrating' (Okubo 2001); these processes, along with intra- and inter-species interactions, induce the formation of spatial patterns for species. A central concept in modelling the dispersal of biological resources is that of *diffusion*. Diffusion is defined as a process whereby the microscopic irregular movement of particles such as cells, bacteria, chemicals, or animals results in some macroscopic regular motion of the group. Biological diffusion is based on random walk models which, when coupled with population growth equations, lead to general reaction-diffusion systems (see, for example, Okubo and Levin 2001; Murray 2003). When only one species is examined, the coupling of classical diffusion with a logistic growth function leads to the so-called Fisher–Kolmogorov equation, which can be written as

$$\frac{\partial x(z,t)}{\partial t} = F(x(z,t)) + D_x \frac{\partial^2 x(z,t)}{\partial z^2} \tag{12}$$

where $x(z,t)$ denotes the concentration of the biomass at spatial point z at time t. The biomass grows according to a standard growth function $F(x)$ which determines the resource's kinetics but also disperses in space with a constant diffusion coefficient D_x. (Nonlinear reaction diffusion equations are associated with propagating wave solutions.) In general, a diffusion process in an ecosystem tends to produce a uniform population density, that is, spatial homogeneity. Thus it might be expected that diffusion would 'stabilize' ecosystems where species disperse and humans intervene through harvesting.

There, is however, one exception, known as 'diffusion induced instability' or 'diffusive instability'. It was Alan Turing (1952) who suggested that under certain conditions reaction-diffusion systems can generate spatially heterogeneous patterns. This is the so-called 'Turing mechanism' for generating diffusion instability. With two interacting species evolving according to

$$\frac{\partial x(z,t)}{\partial t} = F(x,y) + D_x \frac{\partial^2 x(z,t)}{\partial z^2} \tag{13}$$

$$\frac{\partial y(z,t)}{\partial t} = G(x,y) + D_y \frac{\partial^2 y(z,t)}{\partial z^2}, \tag{14}$$

if in the absence of diffusion ($D_x = D_y = 0$) the system tends to a spatially uniform stable steady state, then under certain conditions, depending on the relationship D_x/D_y, spatially heterogeneous patterns can emerge due to diffusion-induced instability.

Spatial variations in ecological systems can also be analysed in terms of meta-population models. A meta-population is a set of local populations occupying isolated patches which are connected by migrating individuals. Meta-population dynamics can be developed for single

or many species (Levin 1974). For the single species case the dynamics become

$$\frac{d\mathbf{x}}{dt} = \mathrm{F}(\mathbf{x})\mathbf{x} + \mathbf{D}\mathbf{x} \tag{15}$$

where $\mathbf{x}=(x_1,\ldots,x_J)$ is a column vector of species densities, $\mathbf{F}$ has its ith row depending on the ith row of $\mathbf{x}$, and $\mathbf{D} = [d_{ij}]$ is a connectivity matrix, where d_{ij} is the rate of movement from patch j to patch i ($j \neq i$). Thus dynamics are local with the exception of movements from one patch to the other.

A more general model encompassing $i = 1,\ldots,n$ species competing for $j = 1,\ldots, J$ limiting resources, with density-dependent growth and interactions across patches $c = 1,\ldots, C$ in a given landscape, can be written as

$$\frac{\dot{x}_{ci}}{x_{ci}} = F_{ic}(\mathbf{x}_c, \mathbf{x}_{-c}) g i_c(\mathbf{R}_c, d_{ic}), \forall i, c \tag{16}$$

$$\dot{R}_{jc} = S_{jc}(\mathbf{R}_c, \mathbf{R}_{-c}) - D_{jc}(\mathbf{x}_c, \mathbf{x}_{-c}, \mathbf{R}_c, \mathbf{R}_{-c}), \forall j, c \tag{17}$$

where $\mathbf{R}_{-c}$, $\mathbf{x}_{-c}$ are respectively vectors of resources and species outside patch c.

(For a detailed analysis, see Brock and Xepapadeas 2002.) A more general setup can be obtained in the context of co-evolutionary models which describe the interactions between population (or biomass) dynamics and mutation (or trait dynamics). Antagonistic co-evolution of species on the one hand and pests or parasites or the other can be described by the so-called Red Queen hypothesis (see, for example, Van Valen 1973, and Kawecki 1998). According to this hypothesis, parasites evolve ceaselessly in response to perpetual evolution of species' (or hosts') resistance. The co-evolution of the parasites' ability to attack (virulence) and the hosts' resistance is expected to indicate persistent fluctuations of resistance and virulence. In this context the Red Queen hypothesis generates a continuous need for variation, resulting in a limit cycle or other non-point attractor in trait space dynamics, which are called Red Queen races. Red Queen cycles are observed in a slow time scale, since trait dynamics are assumed to evolve slowly, in contrast to the population, host-parasite, dynamics which are assumed to evolve fast (see Dieckmann and Law 1996).

A simple co-evolutionary model can be developed in a system with one harvested ('useful') species or host species whose biomass is denoted by x and a parasite denoted by y, where the abundance of x and y depends on the evolution of two characteristics or traits denoted by d and γ (see, for example, the Red Queen dynamic models developed by Krakauer and Jansen 2002), where d affects the fitness of x and γ affects the fitness of y.

Let the growth rates of x and the pathogen y be given by

$$g_x = \frac{\dot{x}}{x} = (s - rx - yQ(d,\gamma)) \tag{18}$$

$$g_y = \frac{\dot{y}}{y} = (xQ(d,\gamma) - \delta). \tag{19}$$

If we measure fitness by growth rates, then $\frac{\partial Q(d,\gamma)}{\partial d} < 0$, so that an increase in d increases fitness of x. In the same way, $\frac{\partial Q(d,\gamma)}{\partial \gamma} > 0$ for an increase in γ to increase fitness of y. In equilibrium of the fast population system where $\dot{x} = \dot{y} = 0$, it holds that

$$\hat{x} = \frac{\delta}{Q(d,\gamma)}, \hat{y} = \frac{s - r\hat{x}}{Q(d,\gamma)}, s \geq r\hat{x}. \tag{20}$$

that On the assumption of constant mutation rates μ_d and μ_γ, the evolutionary dynamics for the traits d and γ, when population dynamics have reached the asymptotically stable steady state, are given by

$$\dot{d} = -\mu_d \hat{x}\hat{y} \frac{\partial Q(d,\gamma)}{\partial d} \tag{21}$$

$$\dot{\gamma} = \mu_\gamma \hat{x}\hat{y} \frac{\partial Q(d,\gamma)}{\partial \gamma}. \tag{22}$$

See Krakauer and Jansen (2002) who, by considering the slow time scale trait dynamics, show that the equilibrium point $(d^*, \gamma^*) : \dot{d} = \dot{\gamma} = 0$ is not attracting; the dynamics spiral away from this

point. This behaviour is the oscillatory, Red Queen dynamics.

Ecological-Economic Modelling

The ecological models developed above are the cornerstones of the development of meaningful ecological-economic models. The impact of humans on the population of species can be realized through direct harvesting h as described in (1) and (2). This type of impact can be easily incorporated into the more general population dynamic models by selecting the harvested species. Human influence can also be realized in an indirect way by having the environmental carrying capacity affected by environmental pollution generated in the non-harvesting sector of the economy, as in (3) and (4), or by having technological considerations affecting catchability coefficients. It is also possible that external environmental conditions which are anthropogenic, such as global warming, can make some parameters associated with population dynamics or mutation dynamics change slowly. This can be modelled in (21) and (22) by considering μ_d and μ_γ as slow varying parameters, defined as $\mu_d(\varepsilon t)$ and $\mu_\gamma(\varepsilon t)$, where $0 < \varepsilon \ll 1$ is the *adiabatic parameter*. This slowly varying system could be used to model slow anthropogenic impacts on ecosystem structure.

However, the size and the severity of the impact of human economies in ecosystems depend on the way in which variables, such as harvesting or other variables which can be chosen by humans (such as emissions, investment in harvesting capacity) and which influence the evolution of ecosystems, are actually chosen. These variables can be regarded as *control variables*, and the way in which they are chosen affects the evolution of ecological variables, such as species biomasses or traits, which can be considered as the *state variables* of the problem.

The typical approach in economics is to associate the choice of the control variables with *optimizing behaviour*. Thus, the control variables are chosen so that a criterion function is optimized, and the economic problem of ecosystem management – where management means choice of control variables – is defined as a formal optimal control problem. In this problem the objective is the optimization of the criterion function subject to the constraints imposed by the structure of the ecosystem. These constraints, which provide the transition equations of the optimal control problem, are the dynamic equations of the ecological models described in the previous section.

The solution of the ecological-economic model, provided it exists, will determine the paths of the state and the control variables and the steady state of the system, which will determine the long-run equilibrium values of the ecological populations as well as the approach dynamics to the steady state. In this context, managed ecological systems which are predominantly nonlinear could exhibit dynamic behaviour characterized by multiple, locally stable and unstable steady states, limit cycles, or the emergence of hysteresis, bifurcations or irreversibilities.

The way in which the objective function is set up and the ecological constraints which are taken into account determine the solution of the ecological-economic model. In principle, a *socially optimal solution* can be distinguished from a *privately optimal solution*. The socially optimal solution corresponds to the so-called problem of the *social planner*, where the objective function takes into account not only benefits from harvesting certain resources of the ecological system, which corresponds to harvesting commercially valuable biomass, but in addition a wide spectrum of flows of services generated by the whole ecosystem. These include, as described above, regulation, cultural or supporting services, existence values, or benefits associated with productivity or insurance gains. If $V(\mathbf{h}(t))$ denotes harvesting benefits at time t associated with harvesting vector $\mathbf{h}$, and $U(\mathbf{x}(t))$ denotes the flow of benefits associated with ecosystem service generated by species biomasses existing in the ecosystem and not removed by harvesting, then the total flow of benefit is $V(\mathbf{h}(t)) + U(\mathbf{x}(t))$. In this formulation, the $V(\,)$ and $U(\,)$ functions are usually assumed to be monotonically increasing and concave. In a more general setup, the total benefit function can be non-separable, defined as $u(\mathbf{h}(t); \mathbf{x}(t))$.

The objective can then be written as

$$\max_{\{\mathbf{h}(t)\}} \int_0^\infty e^{-\rho t}[V(\mathbf{h}(t)) + U(\mathbf{x}(t))]dt \qquad (23)$$

where $\rho \geq 0$ is a discount rate. It should be noted that in principle benefits associated with $V(\mathbf{h}(t))$ can be estimated using market data, while benefits associated with $U(\mathbf{x}(t))$ are hard to estimate because markets for the larger part of the spectrum of ecosystem services are missing. (Valuation of ecosystem services is an open question. For details, see, for example, Bingham et al. 1995.) The social optimum corresponds to the maximization of (23), subject to the constraints imposed by the ecological system. For example, if we use the generalized model of resource competition, the constraints are:

$$\frac{\dot{x}_{ci}}{x_{ci}} = F_{ic}(\mathbf{x}_c, \mathbf{x}_{-c})g_{ic}(\mathbf{R}_c, d_{ic}) - h_{ic}, \forall i, c \qquad (24)$$

$$\dot{R}_{jc} = S_{jc}(\mathbf{R}_c, \mathbf{R}_{-c}) - D_{jc}(\mathbf{x}_c, \mathbf{x}_{-c}, \mathbf{R}_c, \mathbf{R}_{-c}), \forall j, c. \qquad (25)$$

A solution $(\mathbf{h}^*(t), \mathbf{x}^*(t))$ is regarded as the socially optimal solution.

The privately optimal solution is distinguished from the socially optimal by the fact that only harvesting benefits enter the objective function. The assumption is that management is carried out by a 'small' profit-maximizing private agent that ignores the general flows of ecosystem services. In this case, the private agents do not take into account *externalities* associated with their management practices on ecosystem service flow and $U(\mathbf{x}(t)) \equiv 0$. Market externalities associated with the definition of $V(\mathbf{h})$ could relate to imperfections in the markets for the harvested commodities, or to property rights-related externalities, as the well known 'tragedy of the commons' emerging in the harvesting of open access resources.

In general the privately optimal solution $(\mathbf{h}^0(t),\mathbf{x}^0(t))$ will deviate from the socially optimal solution. Another type of externality can be associated with strategic behaviour in resource harvesting if more than one private agent harvests the resource. If $l = 1,\ldots, L$ harvesters are present, then the biomass equation (24) for patch c becomes

$$\frac{\dot{x}_{ci}}{x_{ci}} = F_{ic}(\mathbf{x}_c, \mathbf{x}_{-c})g_{ic}(\mathbf{R}_c, d_{ic}) - \sum_{l=1}^{L} h_{ic}, \forall i, c.$$

In this case the privately optimal solution can be obtained as an *open loop or feedback* Nash equilibrium.

Privately optimal solutions can also be distinguished from the socially optimal by the extent to which the ecological constraints are taken into account. For example, if resource dynamics or trait dynamics are not taken into account in the optimization problem, the management rule will deviate from the social optimum. Furthermore, since *all* the ecological constraints are operating, there will be discrepancies between the perceived evolution of ecosystems under management that ignores certain constraints, and the actual evolution of the ecosystem. Brock and Xepapadeas (2003), show that, by ignoring genetic constraints associated with the development of resistance to genetically modified organisms, the actual system loses any productivity advantage because of resistance development.

These discrepancies might be a cause for *surprises* in ecosystem management. For example, with reference to the co-evolutionary model (18), (19), (20), (21), and (22), profit-maximizing decisions which ignore evolution might steer the system to a certain steady state on a fast time scale, but then the underlying trait dynamics might move the system in slow time to another attractor.

The deviations between the private solution and the social optimum provide a basis for regulation which is similar to the rationale behind the regulation of environmental externalities. Regulation could take the form, in general spatial models of ecosystem management, of species-specific and site-specific taxes on harvesting, or equivalent quota and zoning systems.

See Also

- ▶ Approximate Solutions to Dynamic Models (Linear Methods)
- ▶ Common Property Resources

- ▶ Consumption Externalities
- ▶ Dynamic Programming
- ▶ Environmental Economics
- ▶ Spatial Economics
- ▶ Spatial Econometrics

Bibliography

Bingham, G., et al. 1995. Issues inecosystem valuation: Improving information for decision making. *Ecological Economics* 14(2): 73–90.

Boulding, K. 1965. Earth as a space ship. Washington State University Committee on Space Sciences. Kenneth E. Boulding Papers, Archives (Box # 38), University of Colorado at Boulder Libraries. Online. Available at http://www.colorado.edu/econ/Kenneth.Boulding/spaceship-earth.html. Accessed 13 July 2005.

Brock, W., and A. Xepapadeas. 2002. Optimal ecosystem management when species compete for limiting resources. *Journal of Environmental Economics and Management* 44: 189–230.

Brock, W., and A. Xepapadeas. 2003. Valuing biodiversity from an economic perspective: A unified economic, ecological and genetic approach. *American Economic Review* 93: 1597–1614.

Clark, C. 1990. *Mathematical bioeconomics: The optimal management of renewable resources*. 2nd ed. New York: Wiley.

Dieckmann, U., and R. Law. 1996. The dynamical theory of coevolution: A derivation from stochastic ecological processes. *Journal of Mathematical Biology* 34: 579–612.

Kawecki, T. 1998. Red queen meets Santa Rosalia: Arms races and the evolution of host specialization in organisms with parasitic lifestyles. *American Naturalist* 152(4): 635–651.

Krakauer, D., and V. Jansen. 2002. Red queen dynamics in protein translation. *Journal of Theoretical Biology* 218: 97–109.

Heal, G. 2000. *Nature and the marketplace: Capturing the value of ecosystem services*. Washington, DC: Island Press.

Levin, S. 1974. Dispersion and population interactions. *American Naturalist* 108: 207–228.

Millennium Ecosystem Assessment. 2005. Ecosystems and human well-being, Policy responses. Vol. 3. Washington, DC: Island Press.

Murray, J. 2003. *Mathematical biology*. 3rd ed. Springer: Berlin.

Okubo, A. 2001. Introduction: The mathematics of ecological diffusion. In *Diffusion and ecological problems: Modern perspectives*, 2nd ed., ed. A. Okubo and S. Levin. Berlin: Springer.

Okubo, A., and S. Levin. 2001. The basics of diffusion. In *Diffusion and ecological problems: Modern perspectives*, 2nd ed., ed. A. Okubo and S. Levin. Berlin: Springer.

Roughgarden, J., R. May, and S. Levin. 1989. *Perspectives in ecological theory*. Princeton: Princeton University Press.

Tilman, D. 1982. *Resource competition and community structure*. Princeton: Princeton University Press.

Tilman, D. 1988. *Plant strategies and the dynamics and structure of plant communities*. Princeton: Princeton University Press.

Tilman, D., S. Polasky, and C. Lehman. 2005. Diversity, productivity and temporal stability in the economies of humans and nature. *Journal of Environmental Economics and Management* 49: 405–426.

Tschirhart, J. 2000. General equilibrium of an ecosystem. *Journal of Theoretical Biology* 203: 13–32.

Turing, A. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 237(641): 37–72.

Van Valen, L. 1973. A new evolutionary law. *Evolutionary Theory* 1: 1–30.

Ecological Inference

Gary King, Ori Rosen and Martin Tanner

Abstract

Ecological inference is a general statistical problem where a response variable is not available at the subject level because summary statistics are reported for groups only. It consists of merging information from different databases which are not linked to each other at the record level. We consider an election scenario where in each electoral precinct the fraction of voting-age people who turn out to vote, the fraction of black population and the number of voting-age people are observed. The proportions of blacks and of whites who vote are unobserved because electoral results and census data are not linked.

Keywords

Aggregation; Ecological inference; Likelihood; Markov chain Monte Carlo methods; Method of bounds; Nonparametric models; Statistical approaches

JEL Classifications
C10

The Ecological Inference Problem

For expository purposes, we discuss only an important but simple special case of ecological inference, and adopt the running example and notation from King (1997: ch. 2). The basic problem has two observed variables (T_i and X_i) and two unobserved quantities of interest (β_i^b and β_i^w) for each of p observations. Observations represent aggregate units, such as geographic areas, and each individual-level variable within these units is dichotomous.

To be more specific, in Fig. 1 we observe for each electoral precinct $i (i = 1, \ldots, p)$ the fraction of voting age people who turnout to vote (T_i) and who are black (X_i), along with the number of voting age people (N_i). The quantities of interest, which remain unobserved because of the secret ballot, are the proportions of blacks who vote $\left(\beta_i^b\right)$ and whites who vote $\left(\beta_i^w\right)$. The proportions β_i^b and $\left(\beta_i^w\right)$ are not observed because T_i and X_i are from different data sources (electoral results and census data, respectively) and record linkage is impossible (and illegal), and so the cross-tabulation cannot be computed.

Also of interest are the district-wide fractions of blacks and whites who vote, which are respectively

Race of voting age person	Voting decision: Vote	Voting decision: No Vote	
Black	β_i^b	$1 - \beta_i^b$	X_i
White	β_i^w	$1 - \beta_i^w$	$1 - X_i$
	T_i	$1 - T_i$	

Ecological Inference, Fig. 1 Notation for Precinct i. *Note*: The goal is to estimate the quantities of interest, β_i^b (the fraction of blacks who vote) and β_i^b (the fraction of whites who vote), from the aggregate variables X_i (the fraction of voting age people who are black) and T_i (the fraction of people who vote), along with N_i (the known number of voting age people)

$$B^b = \frac{\sum_{i=1}^{p} N_i X_i \beta_i^b}{\sum_{i=1}^{p} N_i X_i}, \text{ and} \tag{1}$$

$$B^w = \frac{\sum_{i=1}^{p} N_i (1 - X_i) \beta_i^w}{\sum_{i=1}^{p} N_i (1 - X_i)}. \tag{2}$$

These are weighted averages of the corresponding precinct-level quantities. Some methods aim to estimate only B^b and B^w without giving estimates of β_i^b and β_i^w for all i.

Deterministic and Statistical Approaches

The ecological inference literature before King (1997) was bifurcated between supporters of the method of bounds, originally proposed by Duncan and Davis (1953), and supporters of statistical approaches, proposed even before Ogburn and Goltra (1919) but first formalized into a coherent statistical model by Goodman (1953, 1959). (For the historians of science among us: although these two monumental articles were written by two colleagues and friends in the same year and in the same department and university – the Department of Sociology at the University of Chicago – the principal did not discuss their work prior to completion. Even by today's standards, nearly a half century after their publication, the articles are models of clarity and creativity.) Although Goodman and Duncan and Davis moved on to other interests following their seminal contributions, most of the ecological inference literature in the five decades since 1953 was an ongoing war between supporters of these two key approaches, and often without the usual academic decorum.

Extracting Deterministic Information: The Method of Bounds

The purpose of the method of bounds and its generalizations is to extract deterministic information, known with certainty, about the quantities of interest.

The intuition behind these quantities is simple. For example, if a precinct contained 150 African-Americans and 87 people in the precinct voted, then how many of the 150 African-American actually cast their ballot? We do not know exactly, but bounds on the answer are easy to obtain: in this case, the answer must lie between 0 and 87. Indeed, conditional only on the data being correct, [0,87] is a 100 per cent confidence interval. Intervals like this are sometimes narrow enough to draw meaningful inferences, and sometimes they are too wide, but the ability to provide (non-trivial) 100 per cent confidence intervals in even some situations is quite rare in any statistical field.

In general, before any data are seen, the unknown parameters β_i^b and β_i^w are each bounded on the unit interval. Once we observe T_i and X_i they are bounded more narrowly, as:

$$\begin{aligned}\beta_i^b &\in \left[\max\left(0, \frac{T_i-(1-X_i)}{X_i}\right), \min\left(\frac{T_i}{X_i}, 1\right)\right] \\ \beta_i^w &\in \left[\max\left(0, \frac{T_i-X_i}{1-X_i}\right), \min\left(\frac{T_i}{1-X_i}, 1\right)\right].\end{aligned} \tag{3}$$

Deterministic bounds on the district-level quantities B^b and B^w are weighted averages of these precinct-level bounds.

The bounds then indicate that the parameters in each case fall within these deterministic bounds with certainty, and in practice they are almost always narrower than [0,1]. Whether they are narrow enough in any one application depends on the nature of the data.

Extracting Statistical Information: Goodman's Regression

Leo Goodman's (1953, 1959) approach is very different from, but just as important as, Duncan and Davis's. He looked at the same data and focused on the statistical information. His approach examines variation in the marginals (X_i and T_i) over the precincts to attempt to reason back to the district-wide fractions of blacks and whites who vote, $B^b$and B^w. The outlines of this approach and the problems with it have been known at least since Ogburn and Goltra (1919). For example, if in precincts with large proportions of black citizens we observe that many people do not vote, then it may seem reasonable to infer that blacks turn out at lower rates than whites. Indeed, it often is reasonable, but not always. The problem is that it could instead be the case that the whites who happen to live in heavily black precincts are the ones who vote less frequently, yielding the opposite ecological inference to the individual-level truth.

What Goodman accomplished was to formalize the logic of the approach in a simple regression model, and to give the conditions under which estimates from such a model are unbiased. To see this, note first that the accounting identity

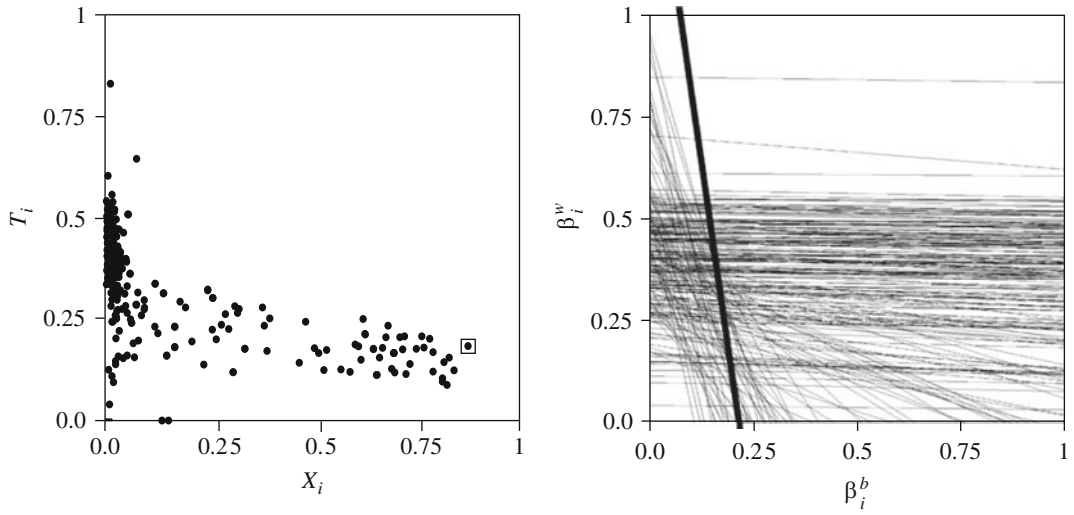
$$T_i = X_i\beta_i^b + (1-X_i)\beta_i^w \tag{4}$$

holds exactly. Then he showed that a regression of T_i on X_i and $(1 - X_i)$ with no constant term could be used to estimate B^b and B^w, respectively. The key assumption necessary for unbiasedness that Goodman identified is that the parameters and X_i be uncorrelated: $\mathrm{Cov}\left(\beta_i^b, X_i\right) = \mathrm{Cov}\left(\beta_i^w, X_i\right) = 0$. In the example, the assumption is that blacks vote in the same proportions in homogeneously black areas as in more integrated areas. Obviously, this is true sometimes and it is false other times. (King 1997: ch. 3, showed that Goodman's assumption was necessary but not sufficient. To have unbiasedness, it must also be true that the parameters and N_i are uncorrelated.)

As Goodman recognized, when this key assumption does not hold, estimates from the model will be biased. Indeed, they can be very biased, outside the deterministic bounds, and even outside the unit interval. This technique has been used extensively since the 1950s, and impossible estimates occur with considerable frequency (some estimates range to a majority of real applications; Achen and Shively 1995).

Extracting Both Deterministic and Statistical Information: King's EI Approach

From 1953 until 1997, the only two approaches used widely in practice were the method of bounds and Goodman's regression. King's



Ecological Inference, Fig. 2 Two views of the same data. *Note*: The left graph is a scatterplot of the observables, X_i by T_i. The right graph displays this same information as a tomography plot of the quantities of interest, β_i^b by β_i^w. Each precinct i that appears as a point in the left graph is a line (rather than a point because of information lost due to aggregation) in the right graph. For example, precinct 52 appears as the dot with a little square around it in the left graph and the dark line in the right graph (*Source*: The data are from King (1997: Figs. 5.1 and 5.5))

(1997) idea was that the insights from these two conflicting literatures in fact do not conflict with each other; the sources of information are largely distinct and can be combined to improve inference overall and synergistically. The idea is to combine the information from the bounds, applied to both quantities of interest for each and every precinct, with a statistical approach for extracting information within the bounds. The amount of information in the bounds depends on the data-set, but for many data-sets it can be considerable. For example, if precincts are spread uniformly over a scatterplot of X_i by T_i, the average bounds on β_i^b and β_i^w are narrowed from [0,1] to less than half of that range – hence eliminating half of the ecological inference problem with certainty. This additional information also helps make the statistical portion of the model far less sensitive to assumptions than previous statistical methods which exclude the information from the bounds.

To illustrate these points, we first present all the information available without making any assumptions, thus extending the bounds approach as far as possible. As a starting point, the left graph in Fig. 2 provides a scatterplot of a sample data set as observed, X_i horizontally by T_i vertically. Each point in this figure corresponds to one precinct, for which we would like to estimate the two unknowns. We display the unknowns in the right graph of the same figure; any point in the right graph portrays values of the two unknowns, β_i^b which is plotted horizontally, and β_i^w which is plotted vertically. Ecological inference involves locating, for each precinct, the one point in this unit square corresponding to the true values of β_i^b and β_i^w, since values outside the square are logically impossible.

To map the knowns onto the unknowns, King begins Goodman's accounting identity from Eq. 4). From this equation, which holds exactly, King solves for one unknown in terms of the other:

$$\beta_i^w = \left(\frac{T_i}{1 - X_i}\right) - \left(\frac{X_i}{1 - X_i}\right)\beta_i^b, \qquad (5)$$

which shows that β_i^w is a *linear* function of β_i^b, where the intercept and slope are known (since they are functions of the data, X_i and T_i).

King then maps the knowns from the left graph onto the right graph by using the linear relationship in Eq. 5). A key point is that each dot on the

left graph can be expressed, without assumptions or loss of information, as what King called a 'tomography' line within the unit square in the right graph. It is precisely the information lost due to aggregation that causes us to have to plot an entire line (on which the true point must fall) rather than the goal of one point for each precinct on the right graph. In fact, the information lost is equivalent to having a graph of the β_i^b by β_i^w points but having the ink smear, making the points into lines and partly but not entirely obscuring the correct positions of the $\left(\beta_i^b, \beta_i^w\right)$ points. (King also showed that the ecological inference problem is mathematically equivalent to the ill-posed 'tomography' problem of many medical imaging procedures, such as CAT and PET scans, where one attempts to reconstruct the inside of an object by passing X-rays through it and gathering information only from the outside. Because the line sketched out by an X-ray is closely analogous to Eq. 5), King labels the latter a *tomography line* and the corresponding graph a *tomography graph*.)

What does a tomography line tell us? Before we know anything, we know that the true $\left(\beta_i^b, \beta_i^w\right)$ point must lie somewhere within the unit square. After X_i and T_i are observed for a precinct, we also know that the true point must fall on a specific line represented by Eq. 5) and appearing in the tomography plot in Fig. 2. In many cases narrowing the region to be searched for the true point from the entire square to the one line in the square can provide a significant amount of information. To see this, consider the point enclosed in a box in the left graph, and the corresponding dark line in the right graph. This precinct, number 52, has observed values of $X_{52} = 0.88$ and $T_{52} = 0.19$. As a result, substituting into Eq. 5) gives $\beta_i^w = 1.58 - 7.33\beta_i^b$, which when plotted appears as the dark line on the right graph. This particular line tells us that, in our search for the true β_{52}^b,β_{52}^w point on the right graph, we can eliminate with certainty all area in the unit square except that on the line, which is clearly an advance over not having the data. Translated into the quantities of interest, this line tells us (by projecting the line downward to the horizontal axis) that, wherever the true point falls on the line, β_{52}^b must fall in the relatively narrow bounds of [0.07,0.21]. Unfortunately, in this case, β_i^w can only be bounded (by projecting to the left) to somewhere within the entire unit interval. More generally, lines that are relatively steep, like this one, tell us a great deal about β_i^b and little about β_i^w. Tomography lines that are relatively flat give narrow bounds on β_{wi} and wide bounds on β_i^b. Lines that cut off the bottom left (or top right) of the figure give narrow bounds on both quantities of interest.

If the only information available to learn about the unknowns in precinct i is X_i and T_i, a tomography line like that in Fig. 2 exhausts all this available information. This line immediately tells us the known bounds on each of the parameters, along with the precise relationship between the two unknowns, but it is not sufficient to narrow in on the right answer any further. Fortunately, additional information exists in the other observations in the same data set (X_j and T_j for all $i \neq j$) which, under the right assumptions, can be used to learn more about β_i^b and β_i^w in our precinct of interest.

In order to borrow statistical strength from all the precincts to learn about β_i^b and β_i^w in precinct i, some assumptions are necessary. The simplest version of King's model (that is, the one most useful for expository purposes) requires three assumptions, each of which can be relaxed in different ways.

First, the set of $\left(\beta_i^b, \beta_i^w\right)$ points must fall in a single cluster within the unit square. The cluster can fall anywhere within the square; it can be widely or narrowly dispersed or highly variable in one unknown and narrow in the other; and the two unknowns can be positively, negatively, or not at all correlated over i. An example that would violate this assumption would be two or more distinct clusters of $\left(\beta_i^b, \beta_i^w\right)$ points, as might result from subsets of observations with fundamentally different data generation processes (such as from markedly different regions). The specific mathematical version of this one-cluster assumption is that β_i^b and β_i^w follow a truncated bivariate normal density

$$\mathrm{TN}\left(\beta_i^b, \beta_i^w \mid \breve{\mathfrak{B}}, \breve{\Sigma}\right) = N\left(\beta_i^b, \beta_i^w \breve{\mathfrak{B}}, \breve{\Sigma}\right) \frac{1\left(\beta_i^b, \beta_i^w\right)}{R\left(\breve{\mathfrak{B}}, \breve{\Sigma}\right)}, \tag{6}$$

where the kernel is the untruncated bivariate normal,

$$\begin{aligned} &\mathrm{N}\left(\beta_i^b, \beta_i^w \breve{\mathfrak{B}}, \breve{\Sigma}\right) \\ &= (2\pi)^{-1} \left|\breve{\Sigma}\right|^{-1/2} \exp\left[-\frac{1}{2}\left(\beta_i - \breve{\mathfrak{B}}\right) \breve{\Sigma}^{-1} \left(\beta_i - \breve{\mathfrak{B}}\right)\right], \end{aligned} \tag{7}$$

and $1\left(\beta_i^b, \beta_i^w\right)$ is an indicator function that equals 1 if $\beta_i^b \in [0,1]$ and $\beta_i^w \in [0,1]$ and zero otherwise. The normalization factor in the denominator, $R\left(\mathfrak{B}^{\cup}, \sum^{\cup}\right)$, is the volume under the untruncated normal distribution above the unit square:

$$R\left(\overset{\cup}{\mathfrak{B}}, \overset{\cup}{\sum}\right) = \int_0^1 \int_0^1 N(\beta^b, \beta^w \mid \overset{\cup}{\mathfrak{B}}, \overset{\cup}{\sum}) d\beta^b d\beta^w \tag{8}$$

When divided into the untruncated normal, this factor keeps the volume under the truncated distribution equal to 1. The parameters of the truncated density, which we summarize as

$$\overset{\cup}{\psi} = \left\{\overset{\cup}{\mathfrak{B}}^b, \overset{\cup}{\mathfrak{B}}^w, \overset{\cup}{\sigma}_b, \overset{\cup}{\sigma}_w, \overset{\cup}{\rho}\right\} = \left\{\overset{\cup}{\mathfrak{B}}, \overset{\cup}{\sum}\right\}, \tag{9}$$

are on the scale of the untruncated normal (and so, for example, $\mathfrak{B}^{b\cup}$ and $\mathfrak{B}^{w\cup}$ need not be constrained to the unit interval even though β_i^b and β_i^w are constrained by this density).

The second assumption, which is necessary to form the likelihood function, is the absence of spatial autocorrelation: conditional on X_i, T_i and T_j are mean independent. Violations of this assumption in empirically reasonable (and even some unreasonable) ways do not seem to induce much if any bias.

The final, and by far the most critical, assumption is that X_i is independent of β_i^b and β_i^w. The three assumptions together produce what has come to be known as King's 'basic' EI model. (The use of EI to name this method comes from the name of his software, available at http://GKing.Harvard.edu) King also generalizes this assumption, in what has come to be known as the 'extended' EI model, by allowing the truncated normal parameters to vary as functions of measured covariates, Z_i^b and Z_i^b, giving:

$$\begin{aligned} \overset{\cup}{\mathfrak{B}} &= \left[\varphi_1 \breve{\sigma}_b^2 + 0.25) + 0.5\right] + \left(Z_i^b - \overline{Z}^b\right) \alpha^b \overset{\cup}{\mathfrak{B}}_i^w \\ &= \left[\varphi_2 \left(\breve{\sigma}_w^2 + 0.25\right) + 0.5\right] + \left(Z_i^w - \overline{Z}^w\right) \alpha^w \end{aligned} \tag{10}$$

where α^b and α^w are parameter vectors to be estimated along with the original model parameters and that have as many elements as Z_i^b and Z_i^w have columns. This relaxes the mean independence assumptions to:

$$\begin{aligned} \mathrm{E}\left(\beta_i^b \mid X_i, Z_i\right) &= \mathrm{E}\left(\beta_i^b \mid Z_i\right) E\left(\beta_i^w \mid X_i, Z_i\right) \\ &= \mathrm{E}\left(\beta_i^w \mid Z_i\right) \end{aligned}$$

Note that this extended model also relaxes the assumptions of truncated bivariate normality, since there is now a separate density being assumed for each observation. Because the bounds, which differ in width and information content for each i, generally provide substantial information, even X_i can be used as a covariate in Z_i. (The recommended default setting in EI includes X_i as a covariate with a prior on its coefficient.) In contrast, under Goodman's regression, which does not include information in the bounds, including X_i leads to an unidentified model (King 1997: sec. 3.2).

These three assumptions – one cluster, no spatial autocorrelation, and mean independence between the regressor and the unknowns conditional on X_i and Z_i – enable one to compute a posterior (or sampling) distribution of the two unknowns in each precinct. A fundamentally important component of EI is that the quantities of interest are not the parameters of the likelihood

but instead come from conditioning on T_i and producing a posterior for β_i^b and β_i^w in each precinct. Failing to condition on T_i and examining the parameters of the truncated bivariate normal only makes sense if the model holds exactly and so is much more model-dependent than King's approach. Since the most important problem in ecological inference modelling is precisely model misspecification, failing to condition on T assumes away the problem without justification. This point is widely regarded as a critical step in applying the EI model (Adolph et al. 2003).

When bounds are narrow, EI model assumptions do not matter much. But, for precincts with wide bounds on a quantity of interest, inferences can become model dependent. This is especially the case with ecological inference problems precisely because of the loss of information due to aggregation. In fact, this loss of information can be expressed by noting that the joint distribution of β_i^b and β_i^w cannot be fully identified from the data without some untestable assumptions. To be precise, distributions with positive mass over *any* curve or combination of curves that connects the bottom left point $\left(\beta_i^b = 0, \beta_i^w = 0\right)$ to the top right point $\left(\beta_i^b = 1, \beta_i^w = 1\right)$ of a tomography plot cannot be rejected by the data (King 1997: 191). Other features of the distribution are estimable. This fundamental indeterminacy is, of course, a problem because it prevents pinning down the quantities of interest with certainty, but it can also be something of an opportunity since different distributional assumptions can lead to the same estimates, especially since only those pieces of the distributions above the tomography lines are used in the final analysis.

Alternative Approaches to Ecological Inference

In the continuing search for more information to bring to bear on ecological inferences, King et al. (1999) extend King's (1997) model another step. They incorporate King's main advance of combining deterministic and statistical information but begin modelling a step earlier at the individuals who make up the counts. They also build a hierarchical Bayesian model, using easily generalizable Markov chain Monte Carlo (MCMC) technology (Tanner 1996).

To define the model formally, let T_i' denote the *number* of voting age people who turn out to vote. At the top level of the hierarchy they assume that T_0' follows a binomial distribution with probability equal to $\theta_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$ and count N_i. Note that at this level it is assumed that the *expectation* of T_i', rather than T_i', is equal to $X_i\beta_i^b + (1 - X_i)\beta_i^w$. In other words, King (1997) models T_i as a continuous proportion, whereas King et al. (1999) recognize the inherently discrete nature of the counts of voters that go into computing this proportion. The two models are connected, of course, since T_i/N_i approaches T_i as N_i gets large.

The connection to King's tomography line can be seen in the contribution of the data from precinct i to the likelihood, which is.

$$\left(X_i\beta_i^b + (1 - X_i\left(\beta_i^w\right)^{T_i'}\right.$$
$$\left.\left(1 - X_i\beta_i^b - (1 - X_i)\beta_i^w\right)^{\left(N_i - T_i'\right)}\right. \tag{11}$$

By taking the logarithm of this contribution to the likelihood and differentiating with respect to $\beta_i^b$and β_i^w , King, Rosen and Tanner show that the maximum of Eq. (11) is not a unique point, but rather a line whose equation is given by the tomography line in Eq. 5). Thus, the log-likelihood for precinct i looks like two playing cards leaning against each other. As long as T_i is fixed and bounded away from 0.5 (and X_i is a fixed known value between 0 and 1), the derivative at this point is seen to increase with N_i, that is, the pitch of the playing cards increases with the sample size. In other words, for large N_i, the log-likelihood for precinct i degenerates from a surface defined over the unit square into a single playing card standing perpendicular to the unit square and oriented along the corresponding tomography line.

At the second level of the hierarchical model, $\beta_i^b$is distributed as a beta density with parameters c_b and d_b and β_i^w follows an independent beta with parameters c_w and d_w. While β_i^b and β_i^w are assumed *a priori* independent, they are *a*

posteriori dependent. At the third and final level of the hierarchical model, the unknown parameters c_b, d_b, c_w and d_w follow an exponential distribution with a large mean.

A key advantage of this model is that it generalizes immediately to arbitrarily large $R \times C$ tables. This approach was pursued by Rosen et al. (2001), who also provided a much faster method of moment-based estimator. For an application, see King et al. (2003).

Wakefield (2004) presents an alternative approach based on the Bayesian paradigm using a Markov chain Monte Carlo inference scheme. King et al. (2004) survey the latest strategies for solving ecological inference problems in various fields, many of which do not fit the textbook case of a 2×2 table with known marginals and unknown cell entries. Staniswalis (2005) proposes a nonparametric model for ecological inference with an application to renal failure data.

Bibliography

Achen, C., and W. Shively. 1995. *Cross-level inference*. Chicago: University of Chicago Press.

Adolph, C., G. King, M. Herron, and K. Shotts. 2003. A consensus position on second stage ecological inference models. *Political Analysis* 11: 86–94.

Duncan, O., and B. Davis. 1953. An alternative to ecological correlation. *American Sociological Review* 18: 665–666.

Goodman, L. 1953. Ecological regressions and the behavior of individuals. *American Sociological Review* 18: 663–666.

Goodman, L. 1959. Some alternatives to ecological correlation. *American Journal of Sociology* 64: 610–624.

King, G. 1997. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton: Princeton University Press.

King, G., O. Rosen, and M. Tanner. 1999. Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research* 28: 61–90.

King, G., O. Rosen, and M. Tanner, eds. 2004. *Ecological inference: New methodological strategies*. Cambridge: Cambridge University Press.

King, G., O. Rosen, M. Tanner, and A. Wagner. 2003. *The ordinary election of Adolf Hitler: A modern voting behavior approach*. Online. Available at https://gking.harvard.edu/files/gking/files/nazivp.pdf Accessed 16 Aug 2006.

Ogburn, W., and I. Goltra. 1919. How women vote: A study of an election in Portland, Oregon. *Political Science Quarterly* 34: 413–433.

Rosen, O., W. Jiang, G. King, and M. Tanner. 2001. Bayesian and frequentist inference for ecological inference: The $R \times C$ case. *Statistica Neerlandica* 55(2): 134–156.

Staniswalis, J. 2005. On fitting generalized non-linear models with varying coefficients. *Computational Statistics and Data Analysis* 50: 893–902.

Tanner, M. 1996. *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. 3rd ed. New York: Springer-Verlag.

Wakefield, J. 2004. Prior and likelihood choices in the analysis of ecological data. In *Ecological inference: New methodological strategies*, ed. G. King, O. Rosen, and M. Tanner. Cambridge: Cambridge University Press.

Econometric Issues in the Presence of Multiple Equilibria

Francesca Molinari

Abstract

Multiplicity of equilibria implies that the relationship between the outcome variable and the exogenous variables characterising a model is a correspondence rather than a function. This results in an incomplete econometric model. Incompleteness complicates identification and statistical inference on functionals of the probability distribution of the population of interest. This is because it implies that the sampling process and the maintained assumptions may be consistent with a set of values for these functionals, rather than with a single one. As a result, the econometric analysis of models with multiple equilibria needs to either: (1) rely on simplifying assumptions that shift focus to outcome features that are common across equilibria; or (2) augment the model with a "selection mechanism" that chooses the equilibrium played in the regions of multiplicity; or (3) maintain only minimal assumptions that partially identify the functionals of interest. Each of these approaches is reviewed, focusing on static game theoretic models.

Keywords

Aumann expectation; Multiple equilibria; Normal form games; Partial identification; Point identification; Qualitative choice models; Random sets

JEL Classifications

C01; C14; C15; C35

The Basic Problem

Finite game theoretic models have been employed to study a wide range of economic decisions, where each agent's utility is allowed to depend on the choice made by each of the other agents. Examples include the analysis of social interaction models (Brock and Durlauf 2001), labour force participation (Bjorn and Vuong 1985), market entry (Bresnahan and Reiss 1988, 1990, 1991; Berry 1992; Bajari et al. 2009; Jia 2008; Ciliberto and Tamer 2009), product differentiation (Mazzeo 2002; Borzekowski and Cohen 2005), advertising (Sweeting 2008), and many others. From the econometric perspective, a finite game is a generalisation of a standard discrete choice model. For example, a bivariate simultaneous response model may give a stochastic representation of equilibria in a two-player, two-action game.

Generically, given a set of payoffs for the agents, finite games may admit multiple equilibria. Multiplicity of equilibria implies that the mapping from the model's exogenous variables to outcomes is a correspondence rather than a function. This violates the classical "principal assumptions" or "coherency conditions" for simultaneous discrete response models discussed extensively in the econometrics literature (e.g. Heckman 1978; Gourieroux et al. 1980; Schmidt 1981; Blundell and Smith 1994; Maddala 1983). Such coherency conditions require the existence of a unique reduced form, mapping the model's exogenous variables and parameters to a unique realisation of the endogenous variable; hence they constrain the model to be recursive or triangular in nature. Tamer (2003), however, clarifies the distinction between a model which is incoherent, so that no reduced form exists, and a model which is incomplete, so that multiple reduced forms may exist. Models with multiple equilibria belong to the latter category. Jovanovic (1989) discusses the observable implications of these models. Berry and Tamer (2007) review and extend a number of results on the identification of entry models extensively used in the empirical literature. The insights in their analysis extend to models where the discrete outcome has larger support.

This article reviews the challenges posed by the presence of multiple equilibria for the econometric analysis of static game theoretic models. These models do not specify how an equilibrium is selected in the regions of the exogenous variables which admit multiple equilibria, and therefore they are "incomplete". Incompleteness complicates identification and statistical inference on functional of the probability distribution of the population of interest, because it implies that the sampling process and the maintained assumptions may be consistent with a set of values for these functionals, rather than with a single one. The literature has provided various approaches to dealing with this basic problem: (1) imposing simplifying assumptions that shift focus to outcome features that are common across equilibria (e.g. Bresnahan and Reiss (1988, 1990, 1991) and Berry (1992), who study entry games where the number, though not the identities, of entrants is uniquely predicted by the model in equilibrium); (2) explicitly modelling a selection mechanism which specifies the equilibrium played in the regions of multiplicity (e.g. Bjorn and Vuong (1985), who choose a constant; and Bajari et al. (2009), who allow for a more flexible, covariate dependent parametrisation); (3) partially identifying and setestimating the parameters, without imposing assumptions on the selection mechanism or on the extent of heterogeneity in payoffs (e.g. Tamer (2003), who also provides large support conditions and exclusion restrictions that guarantee point identification of the payoff parameters, Ciliberto and Tamer (2009), Andrews et al. (2004), Beresteanu et al. (2008)).

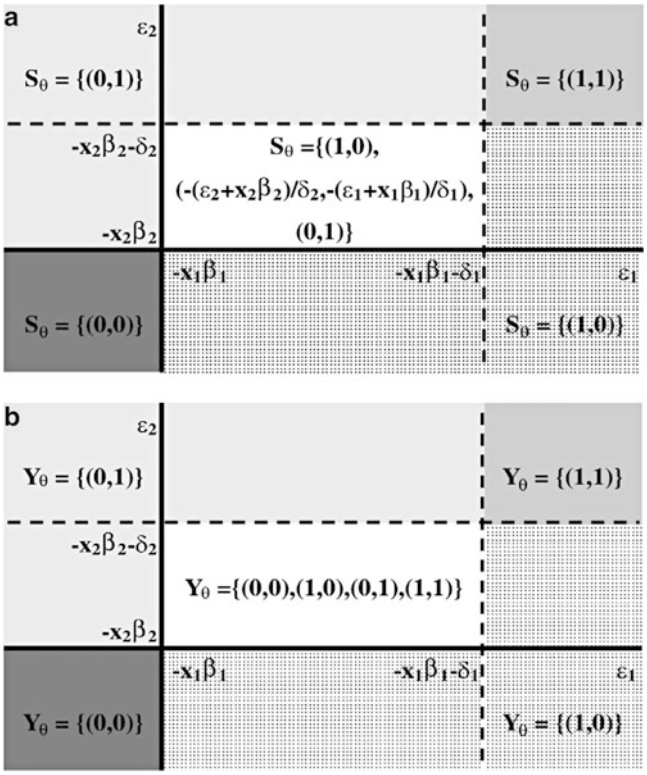
Each of these approaches is reviewed in turn, using the simple example of a complete

information, two-player entry game with multiple mixed strategy Nash equilibria. Similar considerations apply in the econometric analysis of models with more than two players, more than two strategies per player, incomplete information, and/or other solution concepts for the game (e.g. rationalisability, see Aradillas-Lopez and Tamer (2008)). In related models, Brock and Durlauf (2007) and Sweeting (2008) show that the presence of multiple equilibria may actually be beneficial for identification. In this article, however, we do not discuss these cases.

A Simple Example

Consider a static two-player entry game in which players' (stochastic) payoffs are given by $\pi_j = y_j(y_{3-j}\delta_j + x_j\beta_j + \varepsilon_j)$, $j = 1,2$, and mixed strategy Nash equilibrium (MSNE) is the solution concept. Here $y_j \in \{0,1\}$ denotes the action taken by player j, with $y_j = 1$ if player j enters the market and $y_j = 0$ otherwise. Payoff shifters for player j are divided among the ones which are observable both by the players and the econometrician, denoted x_j, and the ones which are observable only by the players, denoted ε_j. For simplicity assume that $(\varepsilon_1,\varepsilon_2)$ is distributed independently of $\underline{x} = (x_1, x_2)$, with a mean-zero normal distribution with covariance matrix Γ. Given the threshold-crossing nature of the model (firms only enter if their profits are positive), let $\gamma_{11} = \gamma_{22} = 1$ and denote the correlation between ε_1 and ε_2 by γ. Let $\sigma_j(\underline{x}, \varepsilon) \in [0, 1]$ denote a mixed strategy for player j, so that she enters the market with probability $\sigma_j(\underline{x}, \varepsilon)$ and stays out of the market with probability $1 - \sigma_j(\underline{x}, \varepsilon)$. The researcher is interested in the parameter vector $\theta = [\delta_1,\delta_2,\beta_1,\beta_2,\gamma] \in \Theta$, with Θ the parameter space. The observable data identify the distribution of equilibrium outcomes and observable payoff shifters, denoted $\mathbf{P}(y, \underline{x})$.

For given $\underline{x}$, Fig. 1a plots the random set of MSNE profiles, denoted $S_\theta(\underline{x}, \varepsilon)$, and Fig. 1b plots the random set of potentially observable MSNE outcomes of the game, denoted $Y_\theta(\underline{x}, \varepsilon)$, as a function of ε_1, ε_2, when $\delta_j < 0$, $j = 1,2$. To



Econometric Issues in the Presence of Multiple Equilibria, Fig. 1 The random set of MSNE profiles S_θ, in panel (**a**), and the random set of potentially observable MSNE outcomes Y_θ, in panel (**b**), in a static, complete information, simultaneous move, two-player entry game with $\delta_j < 0, j = 1, 2$

simplify the notation, in what follows $\mathscr{E}^{t}_{\theta,\underline{x}}$ denotes the region of values for ε where $t \in \{(0,0), (1,0), (0,1), (1,1)\}$ is the unique MSNE of the game; for example, the grey region in Fig. 1a is the region where (0,1) is the unique equilibrium of the game. $\mathscr{E}^{M}_{\theta,\underline{x}}$ denotes the region of values for ε where multiple equilibria occur. In the example, this region is the centre box of Fig. 1a, where (1,0), $\left(\frac{\varepsilon_2+x_2\beta_2}{-\delta_2}, \frac{\varepsilon_1+x_1\beta_1}{-\delta_1}\right)$, (0,1) are the MSNE of the game.

For realisations of $\varepsilon \notin \mathscr{E}^{M}_{\theta,\underline{x}}$, the model admits a unique equilibrium which is in pure strategies, and therefore predicts a unique equilibrium outcome. However, for realisations of $\varepsilon \in \mathscr{E}^{M}_{\theta,\underline{x}}$ the model predicts that any of (0,0), (1,0), (0,1), (1,1) might result as an equilibrium outcome of the game.

One way to reconcile multiplicity of equilibria with the fact that only one equilibrium outcome is realised in each market is to augment the model with an "admissible selection mechanism" $\psi(\cdot; \underline{x}, \varepsilon) : S_\theta(\underline{x}, \varepsilon) \to \Delta^{\kappa-1}$, with $\Delta^{\kappa-1}$ the unit simplex in $\mathscr{R}^{\kappa}$ and κ the cardinality of $S_\theta(\underline{x}, \varepsilon)$. For each $\sigma \in S_\theta(\underline{x}, \varepsilon)$, $\psi(\sigma; \underline{x}, \varepsilon)$ specifies the probability with which that equilibrium is played. (For $\varepsilon \notin \mathscr{E}^{M}_{\theta}$, $S_\theta(\underline{x}, \varepsilon)$ is a singleton and therefore ψ is a scalar identically equal to 1.) For a selection mechanism to be admissible, it is required that $\psi(\sigma; \underline{x}, \varepsilon) \geq 0$ for all $\sigma \in S_\theta(\underline{x}, \varepsilon)$, and $\sum_{\sigma \in S_\theta(\underline{x}, \varepsilon)} \psi(\sigma; \underline{x}, \varepsilon) = 1$. Hence, when no restrictions are placed on it, $\psi(\cdot; \underline{x}, \varepsilon)$ can depend on market unobservables (ε) even after conditioning on market observables ($\underline{x}$). It then follows that given $\underline{x}$, θ, and an admissible ψ, the model predicts that the probability of observing an equilibrium outcome $t \in \{(0,0), (1,0), (0,1), (1,1)\}$ is

$$\mathbf{P}(y = t|\underline{x}; \theta, \psi) = \int \left(\sum_{\sigma \in S_\theta(\underline{x}, \varepsilon)} \psi(\sigma; \underline{x}, \varepsilon) \sigma_1(t_1) \sigma_2(t_2) \right) dF_\theta(\varepsilon).$$

For example, for $t = (0,0)$,

$$\mathbf{P}(y = (0,0)|\underline{x}, \theta, \psi) = \underbrace{\mathbf{P}\left(\varepsilon \in \mathscr{E}^{(0,0)}_{\theta,\underline{x}}\right)}_{\mathbf{P}\left((0,0) \text{ is the unique equilibrium } |\underline{x}\right)} + \underbrace{\int_{\mathscr{E}^{M}_{\theta,\underline{x}}} \left(1 - \frac{\varepsilon_1 + x_1\beta_1}{-\varepsilon_1}\right)\left(1 - \frac{\varepsilon_2 + x_2\beta_2}{-\delta_2}\right)\psi\left(\left(\frac{\varepsilon_2 + x_2\beta_2}{-\delta_2}, \frac{\varepsilon_1 + x_1\beta_1}{-\delta_1}\right), \underline{x}, \varepsilon\right) dF_\theta(\varepsilon)}_{\mathbf{P}\left((0,0) \text{ is observed when multiple equilibria are possible } |\underline{x}\right)}.$$

The identification problem arises because one may find many values for the parameter vector θ which, when coupled with different admissible selection mechanisms ψ, generate the same distribution of outcomes and payoff shifters as the one observed in the data (i.e. $\mathbf{P}(y = t|\underline{x}; \theta, \psi) = \mathbf{P}(y = t|\underline{x})\ \underline{x} - a.s.$).

Point Identification Based on Outcome Features that are Common Across Equilibria

Even in the simple two-player entry game with multiple MSNE described above, multiplicity occurs both in the identity and in the number of players that enter the market in equilibrium. However, if one restricts players to play only pure strategies, and if one assumes $\delta_j < 0$, $j = 1,2$, the model uniquely predicts the equilibrium number of entrants. In other words, there is an outcome feature which is common across equilibria. In this case, under certain restrictions, the "incompleteness" of the model can be circumvented, without the need to introduce a selection mechanism.

Consider first the case that potential entrants are identical in both their observable and unobservable characteristics, so that each firm operating in equilibrium makes the same profit (Bresnahan and Reiss 1991b). With mixed

strategies explicitly ruled out, the equilibrium number of entrants is uniquely predicted by a simple zero profit condition: in equilibrium, no firm will enter if $\delta(1) + x\beta + \varepsilon < 0$, one firm will enter if $\delta(1) + x\beta + \varepsilon \geqslant 0$ and $\delta(2) + x\beta + \varepsilon < 0$, and both firms will enter if $\delta(2) + x\beta + \varepsilon \geqslant 0$. Here $\delta(m)$ denotes the effect on payoffs of m firms entering the market, $m \in \{1,2\}$. Hence point identification of the model's parameters can be achieved, and estimation can be conducted, using familiar techniques for ordered response models. These considerations can be extended to the case where the number of potential entrants is larger, provided that each entrant makes the same profit; see Bresnahan and Reiss (1991b).

Consider now the case that profits differ among firms, as in the example in the previous section (Bresnahan and Reiss 1991a; Berry 1992; Tamer 2003). With heterogeneity in payoffs, but mixed strategies explicitly ruled out, with two players there is still a unique prediction for the number of entrants. Hence the choice probabilities for having an equilibrium with no firms entering the market, and for having an equilibrium with both firms entering the market, are uniquely predicted by the model (this is because (0,0) and (1,1) can only occur as unique equilibrium outcomes of the game), and given by

$$\mathbf{P}(y = (0,0)|\underline{x};\theta) = \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(0,0)}\right),$$
$$\mathbf{P}(y = (1,1)|\underline{x};\theta) = \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(1,1)}\right).$$

In this case, Tamer (2003) shows that under suitable exclusion restrictions and large support conditions on elements of $\underline{x}$, one can use the information in $\mathbf{P}(y = (0, 0)|\underline{x})$ to identify β_1,β_2,γ, and that in $\mathbf{P}(y = (1, 1)|x)$ to identify δ_1, δ_2. A sufficient set of restrictions is as follows. Assume that the matrices x_1 and x_2 have full column rank, and that for either $j = 1$ or $j = 2$, x_j contains an element which is not part of x_{3-j}, has full support, and has a corresponding coefficient in β_j that is nonzero. Then there exist sufficiently large and sufficiently small values of x_j such that player j will always be in or out of the market regardless of what her rival chooses. For these values of x_j, the game simplifies to a single decision problem for player $3 - j$, and one can identify β_{3-j} by using those observations with no entrants and sufficiently small/large values of x_j. One can then learn β_j and γ. Similar reasoning allows one to learn δ_1, δ_2 from $\mathbf{P}(y = (1, 1)|\underline{x})$. Tamer (2003) shows that while using the information contained in the outcomes uniquely predicted by the model suffices for point identification of the parameters of interest, one can obtain efficiency gains by exploiting restrictions on the outcomes of the game resulting from multiple equilibria ((0, 1) and (1, 0)).

Under restrictions on the payoff functions (e.g. homogeneous competition effects) but allowing for a large number of players, Berry (1992) shows that a pure strategy Nash equilibrium for the model exists, and is such that the equilibrium number of entrants is uniquely determined. In this case, inference can be conducted as in nonlinear parametric method of moments problems. Point identification of the model parameters is likely to hold if there is variation in the number of potential entrants across markets; see Berry and Tamer (2007).

Importantly, Tamer (2003) shows that large support and exclusion restrictions can be used to point identify the parameters even when one allows for mixed strategies, and no outcome feature is common across equilibria. Generalisation of this result to games with a larger number of players under related assumptions is discussed in Bajari et al. (2009); see the following section.

Point Identification Based on Specifying a Selection Mechanism

Early on, Bjorn and Vuong (1985) suggested solving the identification problem caused by the presence of multiple equilibria by specifying a selection mechanism that assigns the probability mass of the region of multiplicity among the possible equilibrium outcomes of the game. They considered a two-player, two-action game such as the simple example discussed above, and assumed that players play only pure strategies. (Bjorn and Vuong (1985) were interested in learning the determinants of a husband and wife's

decision to join the labour force. Hence they did not constrain the sign of δ to be known a priori. Here the exposition is simplified by assuming $\delta_j < 0$, $j = 1,2$). For simplicity, they assumed that $\psi((1,0);\underline{x},\varepsilon) = \psi((0,1);\underline{x},\varepsilon) = \frac{1}{2}$ for all $\underline{x}$, ε such that $\varepsilon \in \mathscr{E}^{M}_{\theta,\underline{x}}$. Under this restriction, Bjorn and Vuong (1985) provided a necessary and sufficient condition for the parameters of the model to be point identified, based on the classic criterion of nonsingularity of the information matrix. In their model, estimation can be carried out straightforwardly using maximum likelihood.

Bajari et al. (2009) suggest a more flexible specification of the selection mechanism, while accounting for the possibility that players randomise across their actions. In their model, $\psi(\cdot)$ cannot depend on ($\underline{x}$, ε) directly, but only through players' payoffs. Using this restriction, Bajari et al. (2009) provide a parametrisation of $\psi(\cdot)$ which explicitly accounts for criteria of equilibrium selection often discussed in economic theory. In particular, they assume:

$$\psi(\sigma;\underline{x},\varepsilon) \equiv \psi(\sigma;S_\theta(\underline{x},\varepsilon),\alpha) = \frac{exp(\alpha \cdot z(\sigma,\pi))}{\sum_{\sigma' \in S_\theta(\underline{x},\varepsilon)} exp(\alpha \cdot z(\sigma',\pi))},$$

where z is a vector of covariates including, for example, dummy variables for whether the equilibrium $\sigma \in S_\theta(\underline{x},\varepsilon)$ is in pure strategies, is Pareto dominated, maximises industry profits, and is risk dominant. Bajari et al. (2009) show that under suitable large support conditions on the covariates or with exclusion restrictions, and with scale invariance conditions on $\psi(\cdot)$ (the equilibrium selection probabilities are required to depend only on the relative but not absolute scales of payoffs), both θ and α can be point identified. They then propose a method of simulated moments estimator to estimate these parameters, which embeds a computationally feasible procedure to calculate all the MSNE of the game. Importantly, under the maintained assumptions, this also yields an estimator of the selection mechanism.

Partial Identification of Model Parameters

Given knowledge of $\mathbf{P}(y, \underline{x})$, model parameters can be partially identified even in the absence of assumptions on the nature of competition, heterogeneity of firms, availability of covariates with sufficiently large support and exclusion restrictions, and restrictions on the selection mechanism $\psi(\cdot)$. In particular, the sharp identification region of θ, denoted Θ_I, is given by the set of parameter vectors which are consistent with the sampling process and the maintained modelling assumptions, and therefore may have generated the distribution of observables. Berry and Tamer (2007) provide the following definition of Θ_I in the two-player entry model described in 'A simple example':

$$(0.1) \quad \Theta_I = \left\{ \begin{array}{c} \exists \psi \text{ such that } \forall t \in \{(0,0),(1,0),(0,1),(1,1)\}, \\ \theta \in \Theta : \mathbf{P}(y = t|\underline{x}) = \int \left(\sum_{\sigma \in S_\theta(\underline{x},\varepsilon)} \psi(\sigma;\underline{x},\varepsilon)\sigma_1(t_1)\sigma_2(t_2) \right) dF_\theta(\varepsilon)\underline{x} - a.s. \end{array} \right\}$$

where ψ is an admissible equilibrium selection mechanism as described in 'A simple example'. This formulation is theoretically attractive, but computationally challenging to implement. This is because when no assumptions are placed on it, the selection mechanism ψ may represent an infinite-dimensional nuisance parameter.

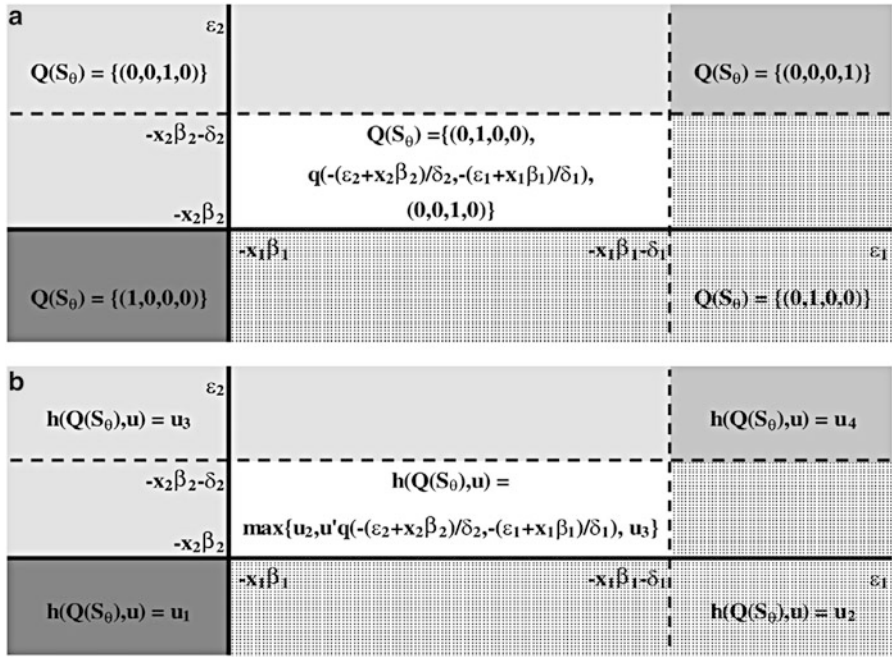
A computationally simple procedure to estimate an outer region for the model parameters is provided by Ciliberto and Tamer (2009). An outer region includes all the parameter values in the parameter space that may have generated the observables, but may include other (infeasible) parameter values as well. They

observe that for a given $t \in \{(0,0), (1,0), (0,1), (1,1)\}$, the model implies that $\mathbf{P}(y = t|\underline{x})$ cannot be larger than the probability that t is *a possible* equilibrium outcome of the game, and cannot be smaller than the probability that t is the *unique* equilibrium outcome of the game. This is because for a given $\theta \in \Theta$ and any realisation of $(\underline{x}, \varepsilon)$ such that t is a possible equilibrium outcome of the game, there can be another outcome $t' \in \{(0,0), (1,0), (0,1), (1,1)\}$ which is also a possible equilibrium outcome of the game, and when both are possible t is selected only part of the time. Similarly, t is certainly realised whenever it is the only possible equilibrium outcome, but it can additionally be realised when it belongs to a set of multiple equilibrium outcomes. In the twoplayer entry game, these considerations yield the following outer region:

$$\Theta_O^{CT} = \left\{ \theta \in \Theta : \begin{array}{c} \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(0,0)}\right) \leq \mathbf{P}(y = (0,0)|\underline{x}) \leq \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(0,0)}\right) \\ + \int_{\mathscr{E}_{\theta,\underline{x}}^{M}} \left(1 - \frac{\varepsilon_1 + x_1\beta_1}{-\delta_1}\right)\left(1 - \frac{\varepsilon_2 + x_2\beta_2}{-\delta_2}\right) dF_\theta(\varepsilon) \\ \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(1,0)}\right) \leq \mathbf{P}(y = (1,0)|\underline{x}) \leq \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(0,0)}\right) \\ + \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{M}\right) \\ \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(0,1)}\right) \leq \mathbf{P}(y = (0,1)|\underline{x}) \leq \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(0,1)}\right) \\ + \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{M}\right) \\ \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(1,1)}\right) \leq \mathbf{P}(y = (1,1)|\underline{x}) \leq \mathbf{P}\left(\varepsilon \in \mathscr{E}_{\theta,\underline{x}}^{(1,1)}\right) \\ + \int_{\mathscr{E}_{\theta,\underline{x}}^{M}} \frac{\varepsilon_1 + x_1\beta_1}{\delta_1} \frac{\varepsilon_2 + x_2\beta_2}{\delta_2} dF_\theta(\varepsilon) \end{array} \right\}$$

Andrews et al. (2004) suggest using only the information provided by the model implication that $\mathbf{P}(y = t|\underline{x})$ cannot be larger than the probability that t is *a possible* equilibrium outcome of the game (hence, using only the upper bounds in the above expression), thereby obtaining an outer region that is simpler to compute than Θ_O^{CT}, but wider.

Exploiting results in Random Set Theory Molchanov (2005), Beresteanu et al. (2008, 2009) propose a formulation of the sharp identification region Θ_I which is computationally tractable. While their formulation is computationally more intensive than Ciliberto and Tamer's, the benefits in terms of identification yielded by their methodology can be substantial; see Beresteanu et al. (2008, 2009) for examples. Their approach can be summarised as follows. Given a $\theta \in \Theta$ and a realisation of $(\underline{x}, \varepsilon)$, one obtains a realisation of the random set of MSNE $S_\theta(\underline{x}, \varepsilon)$; see Fig. 1a. Each of the equilibria in $S_\theta(\underline{x}, \varepsilon)$ determines a probability distribution over the game's outcomes conditional on the realisation of $\underline{x}$ and ε. Denote by $Q(S_\theta(\underline{x}, \varepsilon))$ the random set of such probability distributions; see Fig. 2a. Beresteanu et al. (2008, 2009) establish that the collection of probability distributions over outcomes of the game conditional on $\underline{x}$ which are consistent with the model (i.e. with *all* its implications) is given by the Aumann expectation of $Q(S_\theta(\underline{x}, \varepsilon))$ conditional on $\underline{x}$, denoted $\mathsf{E}(Q(S_\theta(\underline{x}, \varepsilon))|\underline{x})$, which is a closed convex set. (Formally, $\mathsf{E}(Q(S_\theta)|\underline{x}) = \{\mathbf{E}(q|\underline{x}) : q \in Q(S_\theta)$ *a.s.*$\}$, see Molchanov (2005, Definition 2.1.13).) Hence if the model is correctly specified, a candidate value for θ belongs to Θ_I if and only if $\mathbf{P}(y|\underline{x}) \in \mathsf{E}(Q(S_\theta(\underline{x}, \varepsilon))|\underline{x})$, $\underline{x} - a.s.$ In other words, if this condition is satisfied, the candidate θ may have generated the observed conditional distribution $P(y|\underline{x})$. Exploiting the notion of support function of a closed convex set (recall that the support function of a non-empty compact



Econometric Issues in the Presence of Multiple Equilibria, Fig. 2 The random set of probability distributions over outcome profiles implied by MSNE, $Q(S_\theta)$, in panel (**a**), and the support function in direction u of the random set of probability distributions over outcome profiles implied by MSNE, $h(Q(S_\theta),u)$, in panel (**b**), in a static, complete information, simultaneous move, two player entry game with $\delta_j < 0$, $j = 1,2$ and $q(\sigma) = [(1 - \sigma_1)(1 - \sigma_2)\ \sigma_1(1 - \sigma_2)\ (1 - \sigma_1)\sigma_2\ \sigma_1\sigma_2]$

convex set $B \in \mathrm{R}^{ky}$, denoted $h(B,\cdot)$, is given by $h(B, u) = \max_{b \in B} u'b$, $u \in \mathrm{R}^{ky}$), Beresteanu et al. (2008, 2009) show that one can verify this condition by checking if the minimum of a sub-linear (hence convex) function over a convex set is equal to zero. Specifically, they show that

$$\Theta_I = \left\{ \theta \in \Theta : \min_{u:||u||\leq 1.} (\mathbf{E}[h(Q(S_\theta), u)|\underline{x}] - u'\mathbf{P}(y|\underline{x})) = 0 \underline{x} - a.s. \right\},$$

with $h(Q(S_\theta), u)$ the support function of $Q(S_\theta)$ in direction u. This minimisation problem can be solved efficiently using algorithms in convex programming. For certain special cases (e.g., games where players use only pure strategies), Galichon and Henry (2008) provide alternative computational methods based on optimal transportation theory.

Estimation of Θ_I and Θ_O^{CT}, and construction of confidence sets that asymptotically cover these regions with a prespecified probability, can be carried out using the methodology proposed by Chernozhukov et al. (2007).

See Also

- Econometrics
- Identification
- Mixed Strategy Equilibrium
- Partial Identification in Econometrics
- Simulation-Based Estimation

Bibliography

Andrews, D.W.K., S.T. Berry, and P. Jia. 2004. Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location, mimeo.

Aradillas-Lopez, A., and E. Tamer. 2008. The identification power of equilibrium in simple games. *Journal of Business and Economic Statistics* 26(3): 261–310.

Bajari, P., H. Hong, and S. Ryan. 2009. Identification and estimation of a discrete game of complete information. *Econometrica* (forthcoming).

Beresteanu, A., I.S. Molchanov, and F. Molinari. 2008. Sharp identification regions in games. CeMMAP working paper CWP15/08.

Beresteanu, A., I.S.Molchanov, and F. Molinari. 2009. Sharp identification regions in models with convex predictions: Games, individual choice, and incomplete data. CeMMAP working paper CWP27/09.

Berry, S.T. 1992. Estimation of a model of entry in the airline industry. *Econometrica* 60(4): 889–917.

Berry, S.T., and E. Tamer. 2007. Identification in models of oligopoly entry, Chap. 2. In *Advances in economics and econometrics: Theory and application*, Ninth world congress, vol. II, 46–85. Cambridge: Cambridge University Press.

Bjorn, P.A., and Q.H. Vuong. 1985. Simultaneous equations models for dummy endogenous variables: A game theoretic formulation with an application to labor force participation. CalTech DHSS working paper number 557.

Blundell, R., and J.R. Smith. 1994. Coherency and estimation in simultaneous models with censored or qualitative dependent variables. *Journal of Econometrics* 64: 355–373.

Borzekowski, R., and A.M. Cohen. 2005. Estimating strategic complementarities in credit unions' outsourcing decisions. Working paper, Federal Reserve Board of Governors.

Bresnahan, T.F., and P.C. Reiss. 1988. Do entry conditions vary across markets. *Brookings Papers on Economic Activity* 3: 833–871.

Bresnahan, T.F., and P.C. Reiss. 1990. Entry in monopoly markets. *Review of Economic Studies* 57: 531–553.

Bresnahan, T.F., and P.C. Reiss. 1991a. Empirical models of discrete games. *Journal of Econometrics* 48: 57–82.

Bresnahan, T.F., and P.C. Reiss. 1991b. Entry and competition in concentrated markets. *Journal of Political Economy* 99(5): 977–1009.

Brock, W., and S. Durlauf. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68: 235–260.

Brock, W., and S. Durlauf. 2007. Identification of binary choice models with social interactions. *Journal of Econometrics* 140: 52–75.

Chernozhukov, V., H. Hong, and E. Tamer. 2007. Estimation and confidence regions for parameter sets in econometric models. *Econometrica* 75: 1243–1284.

Ciliberto, F., and E. Tamer. 2009. Market structure and multiple equilibria in airline markets. *Econometrica* (forthcoming).

Galichon, A., and M. Henry. 2008. Inference in models with multiple equilibria, mimeo.

Gourieroux, C., J.J. Laffont, and A. Monfort. 1980. Coherency conditions in simultaneous linear equation models with endogenous switching regimes. *Econometrica* 48: 675–695.

Heckman, J. 1978. Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46: 931–959.

Jia, P. 2008. What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica* 76: 1263–1316.

Jovanovic, B. 1989. Observable implications of models with multiple equilibria. *Econometrica* 57: 1431–1437.

Maddala, G.S. 1983. *Limited-dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Mazzeo, M. 2002. Product choice and oligopoly market structure. *RAND Journal of Economics* 33(2): 221–242.

Molchanov, I.S. 2005. *Theory of random sets*. London: Springer.

Schmidt, P. 1981. Constraints on the parameters in simultaneous tobit and probit models, Chap. 12. In *Structural analysis of discrete data and econometric applications*, ed. C. Manski and D. McFadden, 422–434. Cambridge, MA: MIT Press.

Sweeting, A. 2008. The strategic timing of radio commercials: An empirical analysis using multiple equilibria. *RAND Journal of Economics* (forthcoming).

Tamer, E. 2003. Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies* 70: 147–165.

Econometrics

John Geweke, Joel Horowitz and Hashem Pesaran

Abstract

As a unified discipline, econometrics is still relatively young and has been transforming and expanding very rapidly. Major advances have taken place in the analysis of cross-sectional data by means of semiparametric and nonparametric techniques. Heterogeneity of economic relations across individuals, firms and industries is increasingly acknowledged and attempts have been made to take it into account either by integrating out its effects or by modelling the sources of heterogeneity when suitable panel data exist. The counterfactual considerations that underlie policy analysis and treatment valuation have been given a more satisfactory foundation. New time-series

econometric techniques have been developed and employed extensively in the areas of macroeconometrics and finance. Nonlinear econometric techniques are used increasingly in the analysis of cross-section and time-series observations. Applications of Bayesian techniques to econometric problems have been promoted largely by advances in computer power and computational techniques. The use of Bayesian techniques has in turn provided the investigators with a unifying framework where the tasks of forecasting, decision making, model evaluation and learning can be considered as parts of the same interactive and iterative process, thus providing a basis for 'real time econometrics'.

Keywords

Acceptance sampling; Adaptive expectations hypothesis; ARMA processes; Asset pricing models; Asset return volatility; Auctions; Bachelier, L.; Bayesian computation; Bayesian econometrics; Bayesian inference; Benini, R.; Binary logit and probit models; Bootstrap; Building cycle; Bunch maps; Causality in economics and econometrics; Censored regression models; Central limit theorems; Cointegration; Common factors; Conditional hazard functions; Conditional mean functions; Conditional median functions; Confluence analysis; Convexity; Correlation analysis; Cowles Commission; Curse of dimensionality; Davenant, C.; Diagnostic tests; Discrete choice models; Discrete response models; Distributed lags; Douglas, P.H.; Duhem–Quine thesis; Duration models; Dynamic decision models; Dynamic specification; Dynamic stochastic general equilibrium models; Econometric Society; Econometrics; Economic distance; Economic laws; Edgeworth expansions; Edgeworth, F. Y.; Efficient market hypothesis; Engel curve; Error correction models; Euler equations; Experimental economics; Financial econometrics; Fisher, I.; Fisher, R. A.; Fixed effects and random effects; Forecast error variances; Forecast evaluation; Forecasting; Frisch, R. A. K.; Full information maximum likelihood; Galton, F.; Gaussian quadrature; Generalized method of moments; Geometric distributed lag model; Gibbs sampling; Haavelmo, T.; Habit persistence; Hastings–Metropolis algorithm; Hedonic prices; Homogeneity; Hooker, R.H.; Identification; Impulse response analysis; Indirect utility function; Inference; Instrumental variables; Integration; Inventory cycle; Joint hypotheses; Juglar cycle; Juglar, C.; K-class estimators; Kernel estimators; King, G.; Kitchin, J.; Kondratieff, N.; Koopmans, T. C.; Kuznets, S.; Labour market search; Lagrange multiplier; Latent variables; Least absolute deviations estimators; Likelihood ratio; Limited information maximum likelihood; Linear models; Local linear estimation; Logit models; Long waves; Longitudinal data; Lucas critique; Macroeconometric models; Markov chain Monte Carlo methods; Maximum likelihood; Measurement; Measurement errors; Method of simulated moments; Microeconometrics; Microfoundations; Misspecification; Mitchell, W. C.; Model evaluation; Model selection; Model testing; Model uncertainty; Monotonicity; Monte Carlo simulation; Moore, H.L.; Multicollinearity; Multinomial probit model; National Bureau of Economic Research; Nonlinear simultaneous equation models; Nonnested tests; Nonparametric models; Observed variables; Ordinary least squares; Parameter uncertainty; Pearson K.; Petty, W.; Phillips curve; Policy evaluation; Political arithmeticians; Probability; Probability calculus; Probability distribution; Purchasing power parity; Quantile functions; Random assignment; Random utility models; Random variables; Random walk theory; Rational expectations; Real time econometrics; Regional migration; Regression analysis; Revealed preference theory; Saddlepoint expansions; Sampling theory; Schultz, H.; Semiparametric estimation; Sensitivity analysis; Series estimators; Significance tests; Simulated method of moments; Simulation methods; Simultaneous equations models; Simultaneous linear equations; Social experimentation in economics; Spatial econometrics; Specification tests; Splines; Spurious correlation; State dependence; State space models;

Statistical inference; Statistics and economics; Stochastic models; Stock return predictability; Structural change; Structural estimation; Structural VAR; Survival models; Three-stage least squares; Time-series analysis; Tinbergen, J.; Tobit models; Treatment effect; Truncated regression models; Uncovered interest parity; Unit roots; Value distribution; Vector autoregressions (VAR); Vining, R.; Waugh, F.; Weibull hazard model; Working, H.

JEL Classifications
C1

What Is Econometrics?

Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for ex *post* decision/policy evaluation. The term 'econometrics' appears to have been first used by Pawel Ciompa as early as 1910, although it is Ragnar Frisch who takes the credit for coining the term, and for establishing it as a subject in the sense in which it is known today (see Frisch 1936, p. 95; Bjerkholt 1995). By emphasizing the quantitative aspects of economic relationships, econometrics calls for a 'unification' of measurement and theory in economics. Theory without measurement can have only limited relevance for the analysis of actual economic problems; while measurement without theory, being devoid of a framework necessary for the interpretation of the statistical observations, is unlikely to result in a satisfactory explanation of the way economic forces interact with each other. Neither 'theory' nor 'measurement' on its own is sufficient to further our understanding of economic phenomena.

As a unified discipline, econometrics is still relatively young and has been transforming and expanding very rapidly since an earlier version of this article was published in the first edition of *The New Palgrave*: *A Dictionary of Economics* in 1987 (Pesaran 1987a). Major advances have taken place in the analysis of crosssectional data by means of semiparametric and nonparametric techniques. Heterogeneity of economic relations across individuals, firms and industries is increasingly acknowledged, and attempts have been made to take them into account either by integrating out their effects or by modelling the sources of heterogeneity when suitable panel data exists. The counterfactual considerations that underlie policy analysis and treatment evaluation have been given a more satisfactory foundation. New time series econometric techniques have been developed and employed extensively in the areas of macroeconometrics and finance. Nonlinear econometric techniques are used increasingly in the analysis of cross-section and time-series observations. Applications of Bayesian techniques to econometric problems have been given new impetus largely thanks to advances in computer power and computational techniques. The use of Bayesian techniques has in turn provided the investigators with a unifying framework where the tasks of forecasting, decision making, model evaluation and learning can be considered as parts of the same interactive and iterative process; thus paving the way for establishing the foundation of 'real time econometrics'. See Pesaran and Timmermann (2005a).

This article attempts to provide an overview of some of these developments. But to give an idea of the extent to which econometrics has been transformed over the past decades we begin with a brief account of the literature that pre-dates econometrics, and discuss the birth of econometrics and its subsequent developments to the present. Inevitably, our accounts will be brief and non-technical. Readers interested in more details are advised to consultant the specific entries provided in the *New Palgrave* and the excellent general texts by Maddala (2001), Greene (2003), Davidson and MacKinnon (2004), and Wooldridge (2006), as well as texts on specific topics such as Cameron and Trivedi (2005) on microeconometrics, Maddala (1983) on econometric models involving limited-dependent and qualitative variables, Arellano (2003), Baltagi (2005), Hsiao (2003), and Wooldridge (2002) on panel data econometrics, Johansen (1995) on cointegration analysis, Hall (2005) on generalized method of moments, Bauwens et al. (2001), Koop

(2003), Lancaster (2004), and Geweke (2005) on Bayesian econometrics, Bosq (1996), Fan and Gijbels (1996), Horowitz (1998), Härdle (1990), Härdle and Linton (1994), and Pagan and Ullah (1999) on nonparametric and semiparametric econometrics, Campbell et al. (1997) and Gourieroux and Jasiak (2001) on financial econometrics, Granger and Newbold (1986), Lűtkepohl (1991), and Hamilton (1994) on time series analysis.

Quantitative Research in Economics: Historical Backgrounds

Empirical analysis in economics has had a long and fertile history, the origins of which can be traced at least as far back as the work of the 16th-century political arithmeticians such as William Petty, Gregory King and Charles Davenant. The political arithmeticians, led by Sir William Petty, were the first group to make systematic use of facts and figures in their studies. They were primarily interested in the practical issues of their time, ranging from problems of taxation and money to those of international trade and finance. The hallmark of their approach was undoubtedly quantitative, and it was this which distinguished them from their contemporaries. Although the political arithmeticians were primarily and understandably preoccupied with statistical measurement of economic phenomena, the work of Petty, and that of King in particular, represented perhaps the first examples of a unified quantitative–theoretical approach to economics. Indeed Schumpeter in his *History of Economic Analysis* (1954, p. 209) goes as far as to say that the works of the political arithmeticians 'illustrate to perfection, what Econometrics is and what Econometricians are trying to do'.

The first attempt at quantitative economic analysis is attributed to Gregory King, who was the first to fit a linear function of changes in corn prices on deficiencies in the corn harvest, as reported in Charles Davenant (1698). One important consideration in the empirical work of King and others in this early period seems to have been the discovery of 'laws' in economics, very much like those in physics and other natural sciences.

This quest for economic laws was, and to a lesser extent still is, rooted in the desire to give economics the status that Newton had achieved for physics. This was in turn reflected in the conscious adoption of the method of the physical sciences as the dominant mode of empirical enquiry in economics. The Newtonian revolution in physics, and the philosophy of 'physical determinism' that came to be generally accepted in its aftermath, had far-reaching consequences for the method as well as the objectives of research in economics. The uncertain nature of economic relations began to be fully appreciated only with the birth of modern statistics in the late 19th century and as more statistical observations on economic variables started to become available.

The development of statistical theory in the hands of Galton, Edgeworth and Pearson was taken up in economics with speed and diligence. The earliest applications of simple correlation analysis in economics appear to have been carried out by Yule (1895, 1896) on the relationship between pauperism and the method of providing relief, and by Hooker (1901) on the relationship between the marriage rate and the general level of prosperity in the United Kingdom, measured by a variety of economic indicators such as imports, exports, and the movement in corn prices.

Benini (1907), the Italian statistician was the first to make use of the method of multiple regression in economics. But Henry Moore (1914, 1917) was the first to place the statistical estimation of economic relations at the centre of quantitative analysis in economics. Through his relentless efforts, and those of his disciples and followers Paul Douglas, Henry Schultz, Holbrook Working, Fred Waugh and others, Moore in effect laid the foundations of 'statistical economics', the precursor of econometrics. The monumental work of Schultz, *The Theory and the Measurement of Demand* (1938), in the United States and that of Allen and Bowley, *Family Expenditure* (1935), in the United Kingdom, and the pioneering works of Lenoir (1913), Wright (1915, 1928), Working (1927), Tinbergen (1929–1930), and Frisch (1933) on the problem of 'identification'

represented major steps towards this objective. The work of Schultz was exemplary in the way it attempted a unification of theory and measurement in demand analysis; while the work on identification highlighted the importance of 'structural estimation' in econometrics and was a crucial factor in the subsequent developments of econometric methods under the auspices of the Cowles Commission for Research in Economics.

Early empirical research in economics was by no means confined to demand analysis. Louis Bachelier (1900), using time-series data on French equity prices, recognized the random walk character of equity prices, which proved to be the precursor to the vast empirical literature on market efficiency hypothesis that has evolved since the early 1960s. Another important area was research on business cycles, which provided the basis of the later development in time-series analysis and macroeconometric model building and forecasting. Although, through the work of Sir William Petty and other early writers, economists had been aware of the existence of cycles in economic time series, it was not until the early 19th century that the phenomenon of business cycles began to attract the attention that it deserved. Clement Juglar (1819–1905), the French physician turned economist, was the first to make systematic use of time-series data to study business cycles, and is credited with the discovery of an investment cycle of about 7–11 years duration, commonly known as the Juglar cycle. Other economists such as Kitchin, Kuznets and Kondratieff followed Juglar's lead and discovered the inventory cycle (3–5 years duration), the building cycle (15–25 years duration) and the long wave (45–60 years duration), respectively. The emphasis of this early research was on the morphology of cycles and the identification of periodicities. Little attention was paid to the quantification of the relationships that may have underlain the cycles. Indeed, economists working in the National Bureau of Economic Research under the direction of Wesley Mitchell regarded each business cycle as a unique phenomenon and were therefore reluctant to use statistical methods except in a nonparametric manner and for purely descriptive purposes (see, for example, Mitchell 1928; Burns and Mitchell 1947). This view of business cycle research stood in sharp contrast to the econometric approach of Frisch and Tinbergen and culminated in the famous methodological interchange between Tjalling Koopmans and Rutledge Vining about the roles of theory and measurement in applied economics in general and business cycle research in particular. (This interchange appeared in the August 1947 and May 1949 issues of the *Review of Economics and Statistics*.)

The Birth of Econometrics

Although, quantitative economic analysis is a good three centuries old, econometrics as a recognized branch of economics began to emerge only in the 1930s and the 1940s with the foundation of the Econometric Society, the Cowles Commission in the United States, and the Department of Applied Economics (DAE) in Cambridge, England. (An account of the founding of the first two organizations can be found in Christ 1952, 1983, while the history of the DAE is covered in Stone 1978.) This was largely due to the multidisciplinary nature of econometrics, comprising of economic theory, data, econometric methods and computing techniques. Progress in empirical economic analysis often requires synchronous developments in all these four components.

Initially, the emphasis was on the development of econometric methods. The first major debate over econometric method concerned the applicability of the probability calculus and the newly developed sampling theory of R.A. Fisher to the analysis of economic data. Frisch (1934) was highly sceptical of the value of sampling theory and significance tests in econometrics. His objection was not, however, based on the epistemological reasons that lay behind Robbins's and Keynes's criticisms of econometrics. He was more concerned with the problems of multicollinearity and measurement errors which he believed were pervasive in economics; and to deal with the measurement error problem he developed his confluence analysis and the method of 'bunch maps'. Although used by some

econometricians, notably Tinbergen (1939) and Stone (1945), the bunch map analysis did not find much favour with the profession at large. Instead, it was the probabilistic rationalizations of regression analysis, advanced by Koopmans (1937) and Haavelmo (1944), that formed the basis of modern econometrics.

Koopmans did not, however, emphasize the wider issue of the use of stochastic models in econometrics. It was Haavelmo who exploited the idea to the full, and argued for an explicit probability approach to the estimation and testing of economic relations. In his classic paper published as a supplement to *Econometrica* in 1944, Haavelmo defended the probability approach on two grounds. First, he argued that the use of statistical measures such as means, standard errors and correlation coefficients for inferential purposes is justified only if the process generating the data can be cast in terms of a probability model. Second, he argued that the probability approach, far from being limited in its application to economic data, because of its generality is in fact particularly suited for the analysis of 'dependent' and 'nonhomogeneous' observations often encountered in economic research.

The probability model is seen by Haavelmo as a convenient abstraction for the purpose of understanding, or explaining or predicting, events in the real world. But it is not claimed that the model represents reality in all its details. To proceed with quantitative research in any subject, economics included, some degree of formalization is inevitable, and the probability model is one such formalization. The attraction of the probability model as a method of abstraction derives from its generality and flexibility, and the fact that no viable alternative seems to be available. Haavelmo's contribution was also important as it constituted the first systematic defence against Keynes's (1939) influential criticisms of Tinbergen's pioneering research on business cycles and macroeconometric modelling. The objective of Tinbergen's research was twofold: first, to show how a macroeconometric model may be constructed and then used for simulation and policy analysis (Tinbergen 1937); second, 'to submit to statistical test some of the theories which have been put forward regarding the character and causes of cyclical fluctuations in business activity' (Tinbergen 1939, p. 11). Tinbergen assumed a rather limited role for the econometrician in the process of testing economic theories, and argued that it was the responsibility of the 'economist' to specify the theories to be tested. He saw the role of the econometrician as a passive one of estimating the parameters of an economic relation already specified on a priori grounds by an economist. As far as statistical methods were concerned, he employed the regression method and Frisch's method of confluence analysis in a complementary fashion. Although Tinbergen discussed the problems of the determination of time lags, trends, structural stability and the choice of functional forms, he did not propose any systematic methodology for dealing with them. In short, Tinbergen approached the problem of testing theories from a rather weak methodological position. Keynes saw these weaknesses and attacked them with characteristic insight (Keynes 1939). A large part of Keynes's review was in fact concerned with technical difficulties associated with the application of statistical methods to economic data. Apart from the problems of the 'dependent' and 'nonhomogeneous' observations mentioned above, Keynes also emphasized the problems of misspecification, multicollinearity, functional form, dynamic specification, structural stability, and the difficulties associated with the measurement of theoretical variables. By focusing his attack on Tinbergen's attempt at testing economic theories of business cycles, Keynes almost totally ignored the practical significance of Tinbergen's work for econometric model building and policy analysis (for more details, see Pesaran and Smith 1985a).

In his own review of Tinbergen's work, Haavelmo (1943) recognized the main burden of the criticisms of Tinbergen's work by Keynes and others, and argued the need for a general statistical framework to deal with these criticisms. As we have seen, Haavelmo's response, despite the views expressed by Keynes and others, was to rely more, rather than less, on the probability model as the basis of econometric methodology. The technical problems raised by Keynes and

others could now be dealt with in a systematic manner by means of formal probabilistic models. Once the probability model was specified, a solution to the problems of estimation and inference could be obtained by means of either classical or of Bayesian methods. There was little that could now stand in the way of a rapid development of econometric methods.

Early Advances in Econometric Methods

Haavelmo's contribution marked the beginning of a new era in econometrics, and paved the way for the rapid development of econometrics, with the likelihood method gaining importance as a tool for identification, estimation and inference in econometrics.

Identification of Structural Parameters

The first important breakthrough came with a formal solution to the identification problem which had been formulated earlier by Working (1927). By defining the concept of 'structure' in terms of the joint probability distribution of observations, Haavelmo (1944) presented a very general concept of identification and derived the necessary and sufficient conditions for identification of the entire system of equations, including the parameters of the probability distribution of the disturbances. His solution, although general, was rather difficult to apply in practice. Koopmans et al. (1950) used the term 'identification' for the first time in econometrics, and gave the now familiar rank and order conditions for the identification of a single equation in a system of simultaneous *linear* equations. The solution of the identification problem by Koopmans (1949) and Koopmans et al. (1950) was obtained in the case where there are a priori linear restrictions on the structural parameters. They derived rank and order conditions for identifiability of a single equation from a complete system of equations without reference to how the variables of the model are classified as endogenous or exogenous. Other solutions to the identification problem, also allowing for restrictions on the elements of the variance–covariance matrix of the structural disturbances, were later offered by Wegge (1965) and Fisher (1966).

Broadly speaking, a model is said to be identified if all its structural parameters can be obtained from the knowledge of its implied joint probability distribution for the observed variables. In the case of simultaneous equations models prevalent in econometrics, the solution to the identification problem depends on whether there exists a sufficient number of a priori restrictions for the derivation of the structural parameters from the reduced-form parameters. Although the purpose of the model and the focus of the analysis on explaining the variations of some variables in terms of the unexplained variations of other variables is an important consideration, in the final analysis the specification of a minimum number of identifying restrictions was seen by researchers at the Cowles Commission to be the function and the responsibility of 'economic theory'. This attitude was very much reminiscent of the approach adopted earlier by Tinbergen in his business cycle research: the function of economic theory was to provide the specification of the econometric model, and that of econometrics to furnish statistically optimal methods of estimation and inference. More specifically, at the Cowles Commission the primary task of econometrics was seen to be the development of statistically efficient methods for the estimation of structural parameters of an a priori specified system of simultaneous stochastic equations.

More recent developments in identification of structural parameters in context of semiparametric models is discussed below in section "Nonparametric and Semiparametric Estimation". See also Manski (1995).

Estimation and Inference in Simultaneous Equation Models

Initially, under the influence of Haavelmo's contribution, the maximum likelihood (ML) estimation method was emphasized as it yielded consistent estimates. Anderson and Rubin (1949) developed the limited information maximum likelihood (LIML) method, and Koopmans et al. (1950) proposed the full information maximum likelihood (FIML). Both methods are based on the joint probability distribution of the endogenous

variables conditional on the exogenous variables and yield consistent estimates, with the former utilizing all the available a priori restrictions and the latter only those which related to the equation being estimated. Soon, other computationally less demanding estimation methods followed, both for a fully efficient estimation of an entire system of equations and for a consistent estimation of a single equation from a system of equations.

The two-stage least squares (2SLS) procedure was independently proposed by Theil (1954, 1958) and Basmann (1957). At about the same time the instrumental variable (IV) method, which had been developed over a decade earlier by Reiersol (1941, 1945) and Geary (1949) for the estimation of errors-in-variables models, was generalized and applied by Sargan (1958) to the estimation of simultaneous equation models. Sargan's generalized IV estimator (GIVE) provided an asymptotically efficient technique for using surplus instruments in the application of the IV method to econometric problems, and formed the basis of subsequent developments of the generalized method of moments (GMM) estimators introduced subsequently by Hansen (1982). A related class of estimators, known as k-class estimators, was also proposed by Theil (1958). Methods of estimating the entire system of equations which were computationally less demanding than the FIML method were also advanced. These methods also had the advantage that, unlike the FIML, they did not require the full specification of the entire system. These included the three-stage least squares method due to Zellner and Theil (1962), the iterated instrumental variables method based on the work of Lyttkens (1970), Brundy and Jorgenson (1971), and Dhrymes (1971) and the system k-class estimators due to Srivastava (1971) and Savin (1973). Important contributions have also been made in the areas of estimation of simultaneous nonlinear equations (Amemiya 1983), the seemingly unrelated regression equations (SURE) approach proposed by Zellner (1962), and the simultaneous rational expectations models (see section "Model Consistent Expectations" below).

Interest in estimation of simultaneous equation models coincided with the rise of Keynesian economics in early 1960s, and started to wane with the advent of the rational expectations revolution and its emphasis on the GMM estimation of the structural parameters from the Euler equations (first-order optimization conditions). See section "Rational Expectations and the Lucas Critique" below. But, with the rise of the dynamic stochastic general equilibrium models in macroeconometrics, a revival of interest in identification and estimation of nonlinear simultaneous equation models seems quite likely. The recent contribution of Fernandez-Villaverde and Rubio-Ramirez (2005) represents a start in this direction.

Developments in Time Series Econometrics

While the initiative taken at the Cowles Commission led to a rapid expansion of econometric techniques, the application of these techniques to economic problems was rather slow. This was partly due to a lack of adequate computing facilities at the time. A more fundamental reason was the emphasis of the research at the Cowles Commission on the simultaneity problem almost to the exclusion of other econometric problems. Since the early applications of the correlation analysis to economic data by Yule and Hooker, the serial dependence of economic time series and the problem of nonsense or spurious correlation that it could give rise to had been the single most important factor explaining the profession's scepticism concerning the value of regression analysis in economics. A satisfactory solution to the spurious correlation problem was therefore needed before regression analysis of economic time series could be taken seriously. Research on this topic began in the mid-1940s at the Department of Applied Economics (DAE) in Cambridge, England, as a part of a major investigation into the measurement and analysis of consumers' expenditure in the United Kingdom (see Stone et al. 1954). Although the first steps towards the resolution of the spurious correlation problem had been taken by Aitken (1934–1935) and Champernowne (1948), the research in the DAE introduced the problem and its possible solution to the attention of applied economists. Orcutt (1948) studied the autocorrelation pattern of economic time series and showed that most economic time series can be represented

by simple autoregressive processes with similar autoregressive coefficients. Subsequently, Cochrane and Orcutt (1949) made the important point that the major consideration in the analysis of stationary time series was the autocorrelation of the error term in the regression equation and not the autocorrelation of the economic time series themselves. In this way they shifted the focus of attention to the autocorrelation of disturbances as the main source of concern. Although, as it turns out, this is a valid conclusion in the case of regression equations with strictly exogenous regressors, in more realistic set-ups where the regressors are weakly exogenous the serial correlation of the regressors is also likely to be of concern in practice. See, for example, Stambaugh (1999).

Another important and related development was the work of Durbin and Watson (1950, 1951) on the method of testing for residual autocorrelation in the classical regression model. The inferential breakthrough for testing serial correlation in the case of observed time-series data had already been achieved by von Neumann (1941, 1942), and by Hart and von Neumann (1942). The contribution of Durbin and Watson was, however, important from a practical viewpoint as it led to a bounds test for residual autocorrelation which could be applied irrespective of the actual values of the regressors. The independence of the critical bounds of the Durbin–Watson statistic from the matrix of the regressors allowed the application of the statistic as a general diagnostic test, the first of its type in econometrics. The contributions of Cochrane and Orcutt and of Durbin and Watson marked the beginning of a new era in the analysis of economic time-series data and laid down the basis of what is now known as the 'time-series econometrics' approach.

Consolidation and Applications

The work at the Cowles Commission on identification and estimation of the simultaneous equation model and the development of time series techniques paved the way for widespread application of econometric methods to economic and financial problems. This was helped significantly by the rapid expansion of computing facilities, advances in financial and macroeconomic modelling, and the increased availability of economic data-sets, cross section as well as time series.

Macroeconometric Modelling

Inspired by the pioneering work of Tinbergen, Klein (1947, 1950) was the first to construct a macroeconometric model in the tradition of the Cowles Commission. Soon others followed Klein's lead. Over a short space of time macroeconometric models were built for almost every industrialized country, and even for some developing and centrally planned economies. Macroeconometric models became an important tool of ex ante forecasting and economic policy analysis, and started to grow in both size and sophistication. The relatively stable economic environment of the 1950s and 1960s was an important factor in the initial success enjoyed by macroeconometric models. The construction and use of large-scale models presented a number of important computational problems, the solution of which was of fundamental significance, not only for the development of macroeconometric modelling but also for econometric practice in general. In this respect advances in computer technology were clearly instrumental, and without them it is difficult to imagine how the complicated computational problems involved in the estimation and simulation of large-scale models could have been solved. The increasing availability of better and faster computers was also instrumental as far as the types of problems studied and the types of solutions offered in the literature were concerned. For example, recent developments in the area of microeconometrics (see section "Microeconometrics: An Overview" below) could hardly have been possible if it were not for the very important recent advances in computing facilities.

Dynamic Specification

Other areas where econometrics witnessed significant developments included dynamic specification, latent variables, expectations formation, limited dependent variables, discrete choice models, random coefficient models, disequilibrium models, nonlinear estimation, and the

analysis of panel data models. Important advances were also made in the area of Bayesian econometrics, largely thanks to the publication of Zellner's textbook (1971), which built on his earlier work including important papers with George Tiao. The Seminar on Bayesian Inference in Econometrics and Statistics (SBIES) was founded shortly after the publication of the book, and was key in the development and diffusion of Bayesian ideas in econometrics. It was, however, the problem of dynamic specification that initially received the greatest attention. In an important paper, Brown (1952) modelled the hypothesis of habit persistence in consumer behaviour by introducing lagged values of consumption expenditures into an otherwise static Keynesian consumption function. This was a significant step towards the incorporation of dynamics in applied econometric research, and allowed the important distinction to be made between the short-run and the long-run impacts of changes in income on consumption. Soon other researchers followed Brown's lead and employed his autoregressive specification in their empirical work.

The next notable development in the area of dynamic specification was the distributed lag model. Although the idea of distributed lags had been familiar to economists through the pioneering work of Irving Fisher (1930) on the relationship between the nominal interest rate and the expected inflation rate, its application in econometrics was not seriously considered until the mid-1950s. The geometric distributed lag model was used for the first time by Koyck (1954) in a study of investment. Koyck arrived at the geometric distributed lag model via the adaptive expectations hypothesis. This same hypothesis was employed later by Cagan (1956) in a study of demand for money in conditions of hyperinflation, by Friedman (1957) in a study of consumption behaviour and by Nerlove (1958a) in a study of the cobweb phenomenon. The geometric distributed lag model was subsequently generalized by Solow (1960), Jorgenson (1966) and others, and was extensively applied in empirical studies of investment and consumption behaviour. At about the same time Almon (1965) provided a polynomial generalization of I. Fisher's (1937) arithmetic lag distribution which was later extended further by Shiller (1973). Other forms of dynamic specification considered in the literature included the partial adjustment model (Nerlove 1958b; Eisner and Strotz 1963) and the multivariate flexible accelerator model (Treadway 1971) and Sargan's (1964) work on econometric time series analysis which formed the basis of error correction and cointegration analysis that followed next. Following the contributions of Champernowne (1960), Granger and Newbold (1974), and Phillips (1986) the spurious regression problem was better understood, and paved the way for the development of the theory of cointegration. For further details see section "Structural Cointegrating VARs" below.

Techniques for Short-Term Forecasting

Concurrent with the development of dynamic modelling in econometrics there was also a resurgence of interest in time-series methods, used primarily in short-term business forecasting. The dominant work in this field was that of Box and Jenkins (1970), who, building on the pioneering works of Yule (1921, 1926), Slutsky (1927), Wold (1938), Whittle (1963) and others, proposed computationally manageable and asymptotically efficient methods for the estimation and forecasting of univariate autoregressive-moving average (ARMA) processes. Time-series models provided an important and relatively simple benchmark for the evaluation of the forecasting accuracy of econometric models, and further highlighted the significance of dynamic specification in the construction of time-series econometric models. Initially univariate time-series models were viewed as mechanical 'black box' models with little or no basis in economic theory. Their use was seen primarily to be in short-term forecasting. The potential value of modern time-series methods in econometric research was, however, underlined in the work of Cooper (1972) and Nelson (1972) who demonstrated the good forecasting performance of univariate Box–Jenkins models relative to that of large econometric models. These results raised an important question about the adequacy of large econometric models for forecasting as well as for policy analysis. It was argued that a

properly specified structural econometric model should, at least in theory, yield more accurate forecasts than a univariate time-series model. Theoretical justification for this view was provided by Zellner and Palm (1974), followed by Trivedi (1975), Prothero and Wallis (1976), Wallis (1977) and others. These studies showed that Box–Jenkins models could in fact be derived as univariate final form solutions of linear structural econometric models. In theory, the pure time-series model could always be embodied within the structure of an econometric model and in this sense it did not present a 'rival' alternative to econometric modelling. This literature further highlighted the importance of dynamic specification in econometric models and in particular showed that econometric models that are outperformed by simple univariate time-series models most probably suffer from specification errors.

The papers in Elliott et al. (2006) provide excellent reviews of recent developments in economic forecasting techniques.

A New Phase in the Development of Econometrics

With the significant changes taking place in the world economic environment in the 1970s, arising largely from the breakdown of the Bretton Woods system and the quadrupling of oil prices, econometrics entered a new phase of its development. Mainstream macroeconometric models built during the 1950s and 1960s, in an era of relative economic stability with stable energy prices and fixed exchange rates, were no longer capable of adequately capturing the economic realities of the 1970s. As a result, not surprisingly, macroeconometric models and the Keynesian theory that underlay them came under severe attack from theoretical as well as from practical viewpoints. While criticisms of Tinbergen's pioneering attempt at macroeconometric modelling were received with great optimism and led to the development of new and sophisticated estimation techniques and larger and more complicated models, the disenchantment with macroeconometric models in 1970s prompted a much more fundamental reappraisal of quantitative modelling as a tool of forecasting and policy analysis.

At a theoretical level it was argued that econometric relations invariably lack the necessary 'microfoundations', in the sense that they cannot be consistently derived from the optimizing behaviour of economic agents. At a practical level the Cowles Commission approach to the identification and estimation of simultaneous macroeconometric models was questioned by Lucas and Sargent and by Sims, although from different viewpoints (Lucas 1976; Lucas and Sargent 1981; Sims 1980). There was also a move away from macroeconometric models and towards microeconometric research with greater emphasis on matching of econometrics with individual decisions.

It also became increasingly clear that Tinbergen's paradigm where economic relations were taken as given and provided by 'economic theorist' was not adequate. It was rarely the case that economic theory could be relied on for a full specification of the econometric model (Leamer 1978). The emphasis gradually shifted from estimation and inference based on a given tightly parameterized specification to diagnostic testing, specification searches, model uncertainty, model validation, parameter variations, structural breaks, and semiparametric and nonparametric estimation. The choice of approach often governed by the purpose of the investigation, the nature of the economic application, data availability, computing and software technology.

What follows is a brief overview of some of the important developments. Given space limitations there are inevitably significant gaps. These include the important contributions of Granger (1969), Sims (1972), and Engle et al. (1983) on different concepts of 'causality' and 'exogeneity', the literature on disequilibrium models (Quandt 1982; Maddala 1983, 1986), random coefficient models (Swamy 1970; Hsiao and Pesaran 2008, unobserved time series models (Harvey 1989), count regression models (Cameron and Trivedi 1986, 1998), the weak instrument problem (Stock et al. 2002), small sample theory (Phillips 1983; Rothenberg 1984), econometric models of auction pricing (Hendricks and Porter 1988; Laffont et al. 1995).

Rational Expectations and the Lucas Critique

Although the rational expectations hypothesis (REH) was advanced by Muth in 1961, it was not until the early 1970s that it started to have a significant impact on time-series econometrics and on dynamic economic theory in general. What brought the REH into prominence was the work of Lucas (1972, 1973), Sargent (1973), Sargent and Wallace (1975) and others on the new classical explanation of the apparent breakdown of the Phillips curve. The message of the REH for econometrics was clear. By postulating that economic agents form their expectations *endogenously* on the basis of the true model of the economy, and a *correct* understanding of the processes generating exogenous variables of the model, including government policy, the REH raised serious doubts about the invariance of the structural parameters of the mainstream macroeconometric models in the face of changes in government policy. This was highlighted in Lucas's critique of macroeconometric policy evaluation. By means of simple examples Lucas (1976) showed that in models with rational expectations the parameters of the decision rules of economic agents, such as consumption or investment functions, are usually a mixture of the parameters of the agents' objective functions and of the stochastic processes they face as historically given. Therefore, Lucas argued, there is no reason to believe that the 'structure' of the decision rules (or economic relations) would remain invariant under a policy intervention. The implication of the Lucas critique for econometric research was not, however, that policy evaluation could not be done, but rather than the traditional econometric models and methods were not suitable for this purpose. What was required was a separation of the parameters of the policy rule from those of the economic model. Only when these parameters could be identified separately given the knowledge of the joint probability distribution of the variables (both policy and non-policy variables) would it be possible to carry out an econometric analysis of alternative policy options.

There have been a number of reactions to the advent of the rational expectations hypothesis and the Lucas critique that accompanied it.

Model Consistent Expectations

The least controversial reaction has been the adoption of the REH as one of several possible expectations formation hypotheses in an otherwise conventional macroeconometric model containing expectational variables. In this context the REH, by imposing the appropriate cross-equation parametric restrictions, ensures that 'expectations' and 'forecasts' generated by the model are consistent. In this approach the REH is regarded as a convenient and effective method of imposing crossequation parametric restrictions on time series econometric models, and is best viewed as the 'model-consistent' expectations hypothesis. There is now a sizeable literature on solution, identification, and estimation of linear RE models. The canonical form of RE models with forward and backward components is given by

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{B}E\left(\mathbf{y}_{t+1} \middle| F_t\right) + \mathbf{w}_t,$$

where $\mathbf{y}_t$ is a vector of endogenous variables, $E(.|\ F_t)$ is the expectations operator, F_t the publicly available information at time t, and $\mathbf{w}_t$ is a vector of forcing variables. For example, log-linearized version of dynamic general equilibrium models (to be discussed) can all be written as a special case of this equation with plenty of restrictions on the coefficient matrices A and B. In the typical case where $\mathbf{w}_t$ are serially uncorrelated and the solution of the RE model can be assumed to be unique, the RE solution reduces to the vector autoregression (VAR)

$$\mathbf{y}_t = \Phi\mathbf{y}_{t-1} + \mathbf{G}\mathbf{w}_t$$

where Φ and $\mathbf{G}$ are given in terms of the structural parameters:

$$\mathbf{B}\Phi^2 - \Phi + \mathbf{A} = 0, \quad \text{and}\, \mathrm{G} = (\mathbf{I} - \mathbf{B}\Phi)^{-1}.$$

The solution of the RE model can, therefore, be viewed as a restricted form of VAR popularized in econometrics by Sims (1980) as a response in

macroeconometric modelling to the rational expectations revolution. The nature of restrictions is determined by the particular dependence of A and B on a few 'deep' or structural parameters. For general discussion of solution of RE models see, for example, Broze et al. (1985) and Binder and Pesaran (1995). For studies of identification and estimation of linear RE models see, for example, Hansen and Sargent (1980), Wallis (1980), Wickens (1982), and Pesaran (1981, 1987b). These studies show how the standard econometric methods can in principle be adapted to the econometric analysis of rational expectations models.

Detection and Modelling of Structural Breaks

Another reaction to the Lucas critique has been to treat the problem of 'structural change' emphasized by Lucas as one more potential econometric 'problem'. Clements and Hendry (1998, 1999) provide a taxonomy of factors behind structural breaks and forecast failures. Stock and Watson (1996) provide extensive evidence of structural break in macroeconomic time series. It is argued that structural change can result from many factors and need not be associated solely with intended or expected changes in policy. The econometric lesson has been to pay attention to possible breaks in economic relations. There now exists a large body of work on testing for structural change, detection of breaks (single as well as multiple), and modelling of break processes by means of piece-wise linear or non-linear dynamic models (Chow 1960; Brown et al. 1975; Nyblom 1989; Andrews 1993; Andrews and Ploberger 1994; Bai and Perron 1998; Pesaran and Timmermann 2005b, 2007. See also the surveys by Stock 1994; Clements and Hendry 2006). The implications of breaks for short-term and long-term forecasting have also begun to be addressed (McCulloch and Tsay 1993; Koop and Potter 2004a, b; Pesaran et al. 2006).

VAR Macroeconometrics

Unrestricted VARs

The Lucas critique of mainstream macroeconometric modelling also led some econometricians, notably Sims (1980, 1982), to doubt the validity of the Cowles Commission style of achieving identification in econometric models. Sims focused his critique on macroeconometric models with a vector autoregressive (VAR) specification, which was relatively simple to estimate; and its use soon became prevalent in macroeconometric analysis. The view that economic theory cannot be relied on to yield identification of structural models was not new and had been emphasized in the past, for example, by Liu (1960). Sims took this viewpoint a step further and argued that in presence of rational expectations a priori knowledge of lag lengths is indispensable for identification, even when we have distinct strictly exogenous variables shifting supply and demand schedules (Sims 1980, p. 7). While it is true that the REH complicates the necessary conditions for the identification of structural models, the basic issue in the debate over identification still centres on the validity of the classical dichotomy between exogenous and endogenous variables (Pesaran 1981). In the context of closed-economy macroeconometric models where all variables are treated as endogenous, other forms of identification of the structure will be required. Initially, Sims suggested a recursive identification approach where the matrix of contemporaneous effects was assumed to be lower (upper) triangular and the structural shocks orthogonal. Other non-recursive identification schemes soon followed.

Structural VARs

One prominent example was the identification scheme developed in Blanchard and Quah (1989), who distinguished between permanent and transitory shocks and attempted to identify the structural models through long-run restrictions. For example, Blanchard and Quah argued that the effect of a demand shock on real output should be temporary (that is, it should have a zero long-run impact), while a supply shock should have a permanent effect. This approach is known as 'structural VAR' (SVAR) and has been used extensively in the literature. It continues to assume that structural shocks are orthogonal, but uses a mixture of short-run and long-run

restrictions to identify the structural model. In their work Blanchard and Quah considered a bivariate VAR model in real output and unemployment. They assumed real output to be integrated of order 1, or $I(1)$, and viewed unemployment as an $I(0)$, or a stationary variable. This allowed them to associate the shock to one of the equations as permanent, and the shock to the other equation as transitory. In more general settings, such as the one analysed by Gali (1992) and Wickens and Motto (2001), where there are m endogenous variables and r long-run or cointegrating relations, the SVAR approach provides $m(m-r)$ restrictions which are not sufficient to fully identify the model, unless $m=2$ and $r=1$ which is the simple bivariate model considered by Blanchard and Quah (Pagan and Pesaran 2007). In most applications additional short-term restrictions are required. More recently, attempts have also been made to identify structural shocks by means of qualitative restrictions, such as sign restrictions. Notable examples include Canova and de Nicolo (2002), Uhlig (2005), and Peersman (2005).

The focus of the SVAR literature has been on impulse response analysis and forecast error variance decomposition, with the aim of estimating the time profile of the effects of monetary policy, oil price or technology shocks on output and inflation, and deriving the relative importance of these shocks as possible explanations of forecast error variances at different horizons. Typically such analysis is carried out with respect to a single model specification, and at most only parameter uncertainty is taken into account (Kilian 1998). More recently the problem of model uncertainty and its implications for impulse response analysis and forecasting have been recognized. Bayesian and classical approaches to model and parameter uncertainty have been considered. Initially, Bayesian VAR models were developed for use in forecasting as an effective shrinkage procedure in the case of high-dimensional VAR models (Doan et al. 1984; Litterman 1985). The problem of model uncertainty in cointegrating VARs has been addressed in Garratt et al. (2003b, 2006), and Strachan and van Dijk (2006).

Structural Cointegrating VARs

This approach provides the SVAR with the decomposition of shocks into permanent and transitory and gives economic content to the long-run or cointegrating relations that underlie the transitory components. In the simple example of Blanchard and Quah this task is trivially achieved by assuming real output to be I(1) and the unemployment rate to be an I(0) variable. To have shocks with permanent effects some of the variables in the VAR must be non-stationary. This provides a natural link between the SVAR and the unit root and cointegration literature. Identification of the cointegrating relations can be achieved by recourse to economic theory, solvency or arbitrage conditions (Garratt et al. 2003a). Also there are often long-run overidentifying restrictions that can be tested. Once identified and empirically validated, the long-run relations can be embodied within a VAR structure, and the resultant structural vector error correction model identified using theory-based short-run restrictions. The structural shocks can be decomposed into permanent and temporary components using either the multivariate version of the Beveridge and Nelson (1981) decompositions, or the one more recently proposed by Garratt et al. (2006a).

Two or more variables are said to be cointegrated if they are individually integrated (or have a random walk component), but there exists a linear combination of them which is stationary. The concept of cointegration was first introduced by Granger (1986) and more formally developed in Engle and Granger (1987). Rigorous statistical treatments followed in the papers by Johansen (1988, 1991) and Phillips (1991). Many further developments and extensions have taken place with reviews provided in Johansen (1995), Juselius (2006), and Garratt et al. (2006b). The related unit root literature is reviewed by Stock (1994) and Phillips and Xiao (1998).

Macroeconometric Models with Microeconomic Foundations

For policy analysis macroeconometric models need to be based on decisions by individual

households, firms and governments. This is a daunting undertaking and can be achieved only by gross simplification of the complex economic interconnections that exists across millions of decision-makers worldwide. The dynamic stochastic general equilibrium (DSGE) modelling approach attempts to implement this task by focusing on optimal decisions of a few representative agents operating with rational expectations under complete learning. Initially, DSGE models were small and assumed complete markets with instantaneous price adjustments, and as a result did not fit the macroeconomic time series (Kim and Pagan 1995). More recently, Smets and Wouters (2003) have shown that DSGE models with sticky prices and wages along the lines developed by Christiano et al. (2005) are sufficiently rich to match most of the statistical features of the main macroeconomic time series. Moreover, by applying Bayesian estimation techniques, these authors have shown that even relatively large models can be estimated as a system. Bayesian DSGE models have also shown to perform reasonably well in forecasting as compared with standard and Bayesian vector autoregressions. It is also possible to incorporate long-run cointegrating relations within Bayesian DSGE models. The problems of parameter and model uncertainty can also be readily accommodated using data-coherent DSGE models. Other extensions of the DSGE models to allow for learning, regime switches, time variations in shock variances, asset prices, and multi-country interactions are likely to enhance their policy relevance (Del Negro and Schorfheide 2004; Del Negro et al. 2005; An and Schorfheide 2007; Pesaran and Smith 2006). Further progress will also be welcome in the area of macroeconomic policy analysis under model uncertainty, and robust policymaking (Brock and Durlauf 2006; Hansen and Sargent 2007).

Model and Forecast Evaluation

While in the 1950s and 1960s research in econometrics was primarily concerned with the identification and estimation of econometric models, the dissatisfaction with econometrics during the 1970s caused a shift of focus from problems of estimation to those of model evaluation and testing. This shift has been part of a concerted effort to restore confidence in econometrics, and has received attention from Bayesian as well as classical viewpoints. Both these views reject the 'axiom of correct specification' which lies at the basis of most traditional econometric practices, but they differ markedly as how best to proceed.

It is generally agreed, by Bayesians as well as by non-Bayesians, that model evaluation involves considerations other than the examination of the statistical properties of the models, and personal judgements inevitably enter the evaluation process. Models must meet multiple criteria which are often in conflict. They should be relevant in the sense that they ought to be capable of answering the questions for which they are constructed. They should be consistent with the accounting and/or theoretical structure within which they operate. Finally, they should provide adequate representations of the aspects of reality with which they are concerned. These criteria and their interaction are discussed in Pesaran and Smith (1985b). More detailed breakdowns of the criteria of model evaluation can be found in Hendry and Richard (1982) and McAleer et al. (1985). In econometrics it is, however, the criterion of 'adequacy' which is emphasized, often at the expense of relevance and consistency.

The issue of model adequacy in mainstream econometrics is approached either as a model selection problem or as a problem in statistical inference whereby the hypothesis of interest is tested against general or specific alternatives. The use of absolute criteria such as measures of fit/parsimony or formal Bayesian analysis based on posterior odds are notable examples of model selection procedures, while likelihood ratio, Wald and Lagrange multiplier tests of nested hypotheses and Cox's centred log-likelihood ratio tests of non-nested hypotheses are examples of the latter approach. The distinction between these two general approaches basically stems from the way alternative models are treated. In the case of model selection (or model discrimination) all the models under consideration enjoy the same status

and the investigator is not committed a priori to any one of the alternatives. The aim is to choose the model which is likely to perform best with respect to a particular loss function. By contrast, in the hypothesis-testing framework the null hypothesis (or the maintained model) is treated differently from the remaining hypotheses (or models). One important feature of the model-selection strategy is that its application always leads to one model being chosen in preference to other models. But, in the case of hypothesis testing, rejection of all the models under consideration is not ruled out when the models are non-nested. A more detailed discussion of this point is given in Pesaran and Deaton (1978).

Broadly speaking, classical approaches to the problem of model adequacy can be classified depending on how specific the alternative hypotheses are. These are the *general specification tests, the diagnostic tests,* and the *non-nested tests.* The first of these, pioneered by Durbin (1954) and introduced in econometrics by Ramsey (1969), Wu (1973), Hausman (1978), and subsequently developed further by White (1981, 1982) and Hansen (1982), are designed for circumstances where the nature of the alternative hypothesis is kept (sometimes intentionally) rather vague, the purpose being to test the null against a *broad* class of alternatives. (The pioneering contribution of Durbin 1954, in this area has been documented by Nakamura and Nakamura 1981.) Important examples of general specification tests are Ramsey's regression specification error test (RESET) for omitted variables and/or misspecified functional forms, and the Durbin–Hausman–Wu test of misspecification in the context of measurement error models and/or simultaneous equation models. Such general specification tests are particularly useful in the preliminary stages of the modelling exercise.

In the case of diagnostic tests, the model under consideration (viewed as the null hypothesis) is tested against more specific alternatives by embedding it within a general model. Diagnostic tests can then be constructed using the likelihood ratio, Wald or Lagrange multiplier (LM) principles to test for parametric restrictions imposed on the general model. The application of the LM principle to econometric problems is reviewed in the papers by Breusch and Pagan (1980), Godfrey and Wickens (1982), and Engle (1984). An excellent review is provided in Godfrey (1988). Examples of the restrictions that may be of interest as diagnostic checks of model adequacy include zero restrictions, parameter stability, serial correlation, heteroskedasticity, functional forms, and normality of errors. The distinction made here between diagnostic tests and general specification tests is more apparent than real. In practice some diagnostic tests such as tests for serial correlation can also be viewed as a general test of specification. Nevertheless, the distinction helps to focus attention on the purpose behind the tests and the direction along which high power is sought.

The need for non-nested tests arises when the models under consideration belong to separate parametric families in the sense that no single model can be obtained from the others by means of a suitable limiting process. This situation, which is particularly prevalent in econometric research, may arise when models differ with respect to their theoretical underpinnings and/or their auxiliary assumptions. Unlike the general specification tests and diagnostic tests, the application of non-nested tests is appropriate when specific but rival hypotheses for the explanation of the same economic phenomenon have been advanced. Although non-nested tests can also be used as general specification tests, they are designed primarily to have high power against specific models that are seriously entertained in the literature. Building on the pioneering work of Cox (1961, 1962), a number of such tests for single equation models and systems of simultaneous equations have been proposed (Pesaran and Weeks 2001).

The use of statistical tests in econometrics, however, is not a straightforward matter and in most applications does not admit of a clear-cut interpretation. This is especially so in circumstances where test statistics are used not only for checking the adequacy of a *given* model but also as guides to model construction. Such a process of model construction involves specification searches of the type emphasized by Leamer (1978) and presents insurmountable pre-test

problems which in general tend to produce econometric models whose 'adequacy' is more apparent than real. As a result, in evaluating econometric models less reliance should be placed on those indices of model adequacy that are used as guides to model construction, and more emphasis should be given to the performance of models over other data-sets and against rival models.

A closer link between model evaluation and the underlying decision problem is also needed. Granger and Pesaran (2000a, b) discuss this problem in the context of forecast evaluation. A recent survey of forecast evaluation literature can be found in West (2006). Pesaran and Skouras (2002) provide a review from a decision-theoretic perspective.

The subjective Bayesian approach to the treatment of several models begins by assigning a prior probability to each model, with the prior probabilities summing to 1. Since each model is already endowed with a prior probability distribution for its parameters and for the probability distribution of observable data conditional on its parameters, there is then a complete probability distribution over the space of models, parameters, and observable data. (No particular problems arise from non-nesting of models in this framework.) This probability space can then be augmented with the distribution of an object or vector of objects of interest. For example, in a macroeconomic policy setting the models could include VARs, DSGEs and traditional large-scale macroeconomic models, and the vector of interest might include future output growth, interest rates, inflation and unemployment, whose distribution is implied by each of the models considered. Implicit in this formulation is the conditional distribution of the vector of interest conditional on the observed data. Technically, this requires the integration (or marginalization) of parameters in each model as well as the models themselves. As a practical matter this usually proceeds by first computing the probability of each model conditional on the data, and then using these probabilities as weights in averaging the posterior distribution of the vector of interest in each model. It is not necessary to choose one particular model, and indeed to do so would be suboptimal. The ability to actually carry out this simultaneous consideration of multiple models has been enhanced greatly by recent developments in simulation methods, surveyed in section "Integration and Simulation Methods in Econometrics" below; recent texts by Koop (2003), Lancaster (2004), and Geweke (2005) provide technical details. Geweke and Whiteman (2006) specifically outline these methods in the context of economic forecasting.

E

Microeconometrics: An Overview

Partly as a response to the dissatisfaction with macroeconometric time-series research and partly in view of the increasing availability of micro data and computing facilities, since the mid-1980s significant advances have been made in the analysis of micro data. Important micro data-sets have become available on households and firms especially in the United States in such areas as housing, transportation, labour markets and energy. These data sets include various longitudinal surveys (for example, University of Michigan Panel Study of Income Dynamics, and Ohio State National Longitudinal Study Surveys), cross-sectional surveys of family expenditures, population and labour force surveys. This increasing availability of micro-data, while opening up new possibilities for analysis, has also raised a number of new and interesting econometric issues primarily originating from the nature of the data. The errors of measurement are likely to be important in the case of some micro data-sets. The problem of the heterogeneity of economic agents at the micro level cannot be assumed away as readily as is usually done in the case of macro data by appealing to the idea of a 'representative' firm or a 'representative' household.

The nature of micro data, often being qualitative or limited to a particular range of variations, has also called for new econometric models and techniques. Examples include categorical survey responses ('up', 'same' or 'down'), and censored or truncated observations. The models and issues considered in the microeconometric literature are wide ranging and include fixed and random effect

panel data models (for example, Mundlak 1961, 1978), logit and probit models and their multinominal extensions, discrete choice or quantal response models (Manski and McFadden 1981), continuous time duration models (Heckman and Singer 1984), and microeconometric models of count data (Hausman et al. 1984; Cameron and Trivedi 1986).

The fixed or random effect models provide the basic statistical framework and will be discussed in more detailed below. Discrete choice models are based on an explicit characterization of the choice process and arise when individual decision makers are faced with a finite number of alternatives to choose from. Examples of discrete choice models include transportation mode choice (Domenich and McFadden 1975), labour force participation (Heckman and Willis 1977), occupation choice (Boskin 1974), job or firm location (Duncan 1980), and models with neighbourhood effects (Brock and Durlauf 2002). Limited dependent variables models are commonly encountered in the analysis of survey data and are usually categorized into truncated regression models and censored regression models. If all observations on the dependent as well as on the exogenous variables are lost when the dependent variable falls outside a specified range, the model is called *truncated,* and, if only observations on the dependent variable are lost, it is called *censored.* The literature on censored and truncated regression models is vast and overlaps with developments in other disciplines, particularly in biometrics and engineering. Maddala (1983, ch. 6) provides a survey.

The censored regression model was first introduced into economics by Tobin (1958) in his pioneering study of household expenditure on durable goods, where he explicitly allowed for the fact that the dependent variable, namely, the expenditure on durables, cannot be negative. The model suggested by Tobin and its various generalizations are known in economics as Tobit models and are surveyed in detail by Amemiya (1984), and more recently in Cameron and Trivedi (2005, ch. 14). Continuous time duration models, also known as survival models, have been used in analysis of unemployment duration, the period of time spent between jobs, durability of marriage, and so on. Application of survival models to analyse economic data raises a number of important issues resulting primarily from the non-controlled experimental nature of economic observations, limited sample sizes (that is, time periods), and the heterogeneous nature of the economic environment within which agents operate. These issues are clearly not confined to duration models and are also present in the case of other microeconometric investigations that are based on time series or cross-section or panel data.

Partly in response to the uncertainties inherent in econometric results based on non-experimental data, there has also been a significant move towards social experimentation, and experimental economics in general. A social experiment aims at isolating the effects of a policy change (or a treatment effect) by comparing the consequences of an exogenous variation in the economic environment of a set of experimental subjects known as the 'treatment' group with those of a 'control' group that have not been subject to the change. The basic idea goes back to the early work of R.A. Fisher (1928) on randomized trials, and has been applied extensively in agricultural and biomedical research. The case for social experimentation in economics is discussed in Burtless (1995). Hausman and Wise (1985) and Heckman and Smith (1995) consider a number of actual social experiments carried out in the United States, and discuss their scope and limitations.

Experimental economics tries to avoid some of the limitations of working with observations obtained from natural or social experiments by using data from laboratory experiments to test economic theories by fixing some of the factors and identifying the effects of other factors in a way that allows *ceteris paribus* comparisons. A wide range of topics and issues are covered in this literature, such as individual choice behaviour, bargaining, provision of public goods, theories of learning, auction markets, and behavioural finance. A comprehensive review of major areas of experimental research in economics is provided in Kagel and Roth (1995).

These developments have posed new problems and challenges in the areas of experimental

design, statistical methods and policy analysis. Another important aspect of recent developments in microeconometric literature relates to the use of microanalytic simulation models for policy analysis and evaluation to reform packages in areas such as health care, taxation, social security systems, and transportation networks. Cameron and Trivedi (2005) review the recent developments in methods and application of microeconometrics. Some of these topics will be discussed in more detail below.

Econometrics of Panel Data

Panel data models are used in many areas of econometrics, although initially they were developed primarily for the analysis of micro behaviour, and focused on panels formed from cross-section of N individual households or firms surveyed for T successive time periods. These types of panels are often referred to as 'micropanels'. In social and behavioural sciences they are also known as longitudinal data or panels. The literature on micro-panels typically takes N to be quite large (in hundreds) and T rather small, often less than ten. But more recently, with the increasing availability of financial and macroeconomic data, analyses of panels where both N and T are relatively large have also been considered. Examples of such data-sets include time series of company data from Datastream, country data from International Financial Statistics or the Penn World Table, and county and state data from national statistical offices. There are also pseudo panels of firms and consumers composed of repeated cross sections that cover cross-section units that are not necessarily identical but are observed over relatively long time periods. Since the available cross-section observations do not (necessarily) relate to the same individual unit, some form of grouping of the cross-section units is needed. Once the grouping criteria are set, the estimation can proceed using fixed effects estimation applied to group averages if the number of observations per group is sufficiently large; otherwise possible measurement errors of the group averages also need to be taken into account. Deaton (1985) pioneered the econometric analysis of pseudo panels. Verbeek (2008) provides a recent review.

Use of panels can enhance the power of empirical analysis and allows estimation of parameters that might not have been identified using the time or the cross-section dimensions alone. These benefits come at a cost. In the case of linear panel data models with a short time span the increased power is usually achieved under assumptions of parameter homogeneity and error cross-section independence. Short panels with autocorrelated disturbances also pose a new identification problem, namely, how to distinguished between dynamics and state dependence (Arellano 2003, ch. 5). In panels with fixed effects the homogeneity assumption is relaxed somewhat by allowing the intercepts in the panel regressions to vary freely over the cross-section units, but continues to maintain the error cross-section independence assumption. The random coefficient specification of Swamy (1970) further relaxes the slope homogeneity assumption, and represents an important generalization of the random effects model (Hsiao and Pesaran 2007). In micro-panels where T is small cross-section dependence can be dealt with if it can be attributed to spatial (economic or geographic) effects. Anselin (1988) and Anselin et al. (2007) provide surveys of the literature on spatial econometrics. A number of studies have also used measures such as trade or capital flows to capture economic distance, as in Conley and Topa (2002), Conley and Dupor (2003), and Pesaran et al. (2004).

Allowing for dynamics in panels with fixed effects also presents additional difficulties; for example, the standard within-group estimator will be inconsistent unless $T \to \infty$ (Nickell 1981). In linear dynamic panels the incidental parameter problem (the unobserved heterogeneity) can be resolved by first differencing the model and then estimating the resultant first-differenced specification by instrumental variables or by the method of transformed likelihood (Anderson and Hsiao 1981, 1982; Holtz-Eakin et al. 1988; Arellano and Bond 1991; Hsiao et al. 2002). A similar procedure can also be followed in the case of short T panel VARs (Binder et al. 2005). But other approaches are needed for nonlinear

panel data models. See, for example, Honoré and Kyriazidou (2000) and review of the literature on nonlinear panels in Arellano and Honoré (2001). Relaxing the assumption of slope homogeneity in dynamic panels is also problematic, and neglecting to take account of slope heterogeneity will lead to inconsistent estimators. In the presence of slope heterogeneity Pesaran and Smith (1995) show that the within-group estimator remains inconsistent even if both N and $T \to \infty$. A Bayesian approach to estimation of micro dynamic panels with random slope coefficients is proposed in Hsiao et al. (1999).

To deal with general dynamic specifications, possible slope heterogeneity and error cross-section dependence, large T and N panels are required. In the case of such large panels it is possible to allow for richer dynamics and parameter heterogeneity. Cross-section dependence of errors can also be dealt with using residual common factor structures. These extensions are particularly relevant to the analysis of purchasing power parity hypothesis (O'Connell 1998; Imbs et al. 2005; Pedroni 2001; Smith et al. 2004), output convergence (Durlauf et al. 2005; Pesaran 2007b), the Fisher effect (Westerlund 2005), house price convergence (Holly et al. 2006), regional migration (Fachin 2006), and uncovered interest parity (Moon and Perron 2007). The econometric methods developed for large panels has to take into account the relationship between the increasing number of time periods and cross-section units (Phillips and Moon 1999). The relative expansion rates of N and T could have important consequences for the asymptotic and small sample properties of the panel estimators and tests. This is because fixed T estimation bias tend to magnify with increases in the cross-section dimension, and it is important that any bias in the T dimension is corrected in such a way that its overall impact disappears as both N and $T \to \infty$, jointly.

The first generation panel unit root tests proposed, for example, by Levin et al. (2002) and Im et al. (2003) allowed for parameter heterogeneity but assumed errors were cross-sectionally independent. More recently, panel unit root tests that allow for error cross-section dependence have been proposed by Bai and Ng (2004), Moon and Perron (2004), and Pesaran (2007a). As compared with panel unit root tests, the analysis of cointegration in panels is still at an early stage of its development. So far the focus of the panel cointegration literature has been on residual-based approaches, although there has been a number of attempts at the development of system approaches as well (Pedroni 2004). But once cointegration is established the long-run parameters can be estimated efficiently using techniques similar to the ones proposed in the case of single time-series models. These estimation techniques can also be modified to allow for error cross-section dependence (Pesaran 2007a). Surveys of the panel unit root and cointegration literature are provided by Banerjee (1999), Baltagi and Kao (2000), Choi (2006), and Breitung and Pesaran (2008).

The micro and macro panel literature is vast and growing. For the analysis of many economic problems, further progress is needed in the analysis of nonlinear panels, testing and modelling of error cross-section dependence, dynamics, and neglected heterogeneity. For general reviews of panel data econometrics, see Arellano (2003), Baltagi (2005), Hsiao (2003), and Wooldridge (2002).

Nonparametric and Semiparametric Estimation

Much empirical research is concerned with estimating conditional mean, median, or hazard functions. For example, a wage equation gives the mean, median or, possibly, some other quantile of wages of employed individuals conditional on characteristics such as years of work experience and education. A hedonic price function gives the mean price of a good conditional on its characteristics. The function of interest is rarely known a priori and must be estimated from data on the relevant variables. For example, a wage equation is estimated from data on the wages, experience, education and, possibly, other characteristics of individuals. Economic theory rarely gives useful

guidance on the form (or shape) of a conditional mean, median, or hazard function. Consequently, the form of the function must either be assumed or inferred through the estimation procedure.

The most frequently used estimation methods assume that the function of interest is known up to a set of constant parameters that can be estimated from data. Models in which the only unknown quantities are a finite set of constant parameters are called 'parametric'. A linear model that is estimated by ordinary least squares is a familiar and frequently used example of a parametric model. Indeed, linear models and ordinary least squares have been the workhorses of applied econometrics since its inception. It is not difficult to see why. Linear models and ordinary least squares are easy to work with both analytically and computationally, and the estimation results are easy to interpret. Other examples of widely used parametric models are binary logit and probit models if the dependent variable is binary (for example, an indicator of whether an individual is employed or whether a commuter uses automobile or public transit for a trip to work) and the Weibull hazard model if the dependent variable is a duration (for example, the duration of a spell of employment or unemployment).

Although parametric models are easy to work with, they are rarely justified by theoretical or other a priori considerations and often fit the available data badly. Horowitz (2001), Horowitz and Savin (2001), Horowitz and Lee (2002), and Pagan and Ullah (1999) provide examples. The examples also show that conclusions drawn from a convenient but incorrectly specified model can be very misleading. Of course, applied econometricians are aware of the problem of specification error. Many investigators attempt to deal with it by carrying out a specification search in which several different models are estimated and conclusions are based on the one that appears to fit the data best. Specification searches may be unavoidable in some applications, but they have many undesirable properties. There is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investigator's model selection criteria. Moreover, the search process invalidates the statistical theory on which inference is based.

Given this situation, it is reasonable to ask whether conditional mean and other functions of interest in applications can be estimated nonparametrically, that is, without making a priori assumptions about their functional forms. The answer is clearly 'yes' in a model whose explanatory variables are all discrete. If the explanatory variables are discrete, then each set of values of these variables defines a data cell. One can estimate the conditional mean of the dependent variable by averaging its values within each cell. Similarly, one can estimate the conditional median cell by cell.

If the explanatory variables are continuous, they cannot be grouped into cells. Nonetheless, it is possible to estimate conditional mean and median functions that satisfy mild smoothness conditions without making a priori assumptions about their shapes. Techniques for doing this have been developed mainly in statistics, beginning with Nadaraya's (1964) and Watson's (1964) nonparametric estimator of a conditional mean function. The Nadaraya–Watson estimator, which is also called a kernel estimator, is a weighted average of the observed values of the dependent variable. More specifically, suppose that the dependent variable is Y, the explanatory variable is X, and the data consist of observations $\{Y_i, X_i : i = 1, \ldots, n\}$. Then the Nadaraya–Watson estimator of the mean of Y at $X = x$ is a weighted average of the Y_i's. Y_i's corresponding to X_i's that are close to x get more weight than do Y_i's corresponding to X_i's that are far from x. The statistical properties of the Nadaraya–Watson estimator have been extensively investigated for both crosssectional and time-series data, and the estimator has been widely used in applications. For example, Blundell et al. (2003) used kernel estimates of Engel curves in an investigation of the consistency of household-level data and revealed preference theory. Hausman and Newey (1995) used kernel estimates of demand functions to estimate the equivalent variation for changes in

gasoline prices and the deadweight losses associated with increases in gasoline taxes. Kernel-based methods have also been developed for estimating conditional quantile and hazard functions.

There are other important nonparametric methods for estimating conditional mean functions. Local linear estimation and series or sieve estimation are especially useful in applications. Local linear estimation consists of estimating the mean of Y at $X = x$ by using a form of weighted least squares to fit a linear model to the data. The weights are such that observations (Y_i, X_i) for which X_i is close to x receive more weight than do observations for which X_i is far from x. In comparison with the Nadaraya–Watson estimator, local linear estimation has important advantages relating to bias and behaviour near the boundaries of the data. These are discussed in the book by Fan and Gijbels (1996), among other places.

A series estimator begins by expressing the true conditional mean (or quantile) function as an infinite series expansion using basis functions such as sines and cosines, orthogonal polynomials, or splines. The coefficients of a truncated version of the series are then estimated by ordinary least squares. The statistical properties of series estimators are described by Newey (1997). Hausman and Newey (1995) give an example of their use in an economic application.

Nonparametric models and estimates essentially eliminate the possibility of misspecification of a conditional mean or quantile function (that is, they consistently estimate the true function), but they have important disadvantages that limit their usefulness in applied econometrics. One important problem is that the precision of a nonparametric estimator decreases rapidly as the dimension of the explanatory variable X increases. This phenomenon is called the 'curse of dimensionality'. It can be understood most easily by considering the case in which the explanatory variables are all discrete. Suppose the data contain 500 observations of Y and X. Suppose, further, that X is a K-component vector and that each component can take five different values. Then the values of X generate 5^k cells. If $K = 4$, which is not unusual in applied econometrics, then there are 625 cells, or more cells than observations. Thus, estimates of the conditional mean function are likely to be very imprecise for most cells because they will contain few observations. Moreover, there will be at least 125 cells that contain no data and, consequently, for which the conditional mean function cannot be estimated at all. It has been proved that the curse of dimensionality is unavoidable in nonparametric estimation. As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if X is multidimensional.

Another problem is that nonparametric estimates can be difficult to display, communicate, and interpret when X is multidimensional. Nonparametric estimates do not have simple analytic forms. If X is one-or two-dimensional, then the estimate of the function of interest can be displayed graphically, but only reduced-dimension projections can be displayed when X has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of an estimate.

A further problem with nonparametric estimation is that it does not permit extrapolation. For example, in the case of a conditional mean function it does not provide predictions of the mean of Y at values of x that are outside of the range of the data on X. This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, in nonparametric estimation it can be difficult to impose restrictions suggested by economic or other theory. Matzkin (1994) discusses this issue.

The problems of nonparametric estimation have led to the development of so-called semiparametric methods that offer a compromise between parametric and nonparametric estimation. Semiparametric methods make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. Semiparametric methods permit greater estimation precision than do nonparametric methods when X is multidimensional. Semiparametric estimation results are usually

easier to display and interpret than are nonparametric ones, and provide limited capabilities for extrapolation.

In econometrics, semiparametric estimation began with Manski's (1975, 1985) and Cosslett's (1983) work on estimating discrete-choice random-utility models. McFadden had introduced multinomial logit random utility models. These models assume that the random components of the utility function are independently and identically distributed with the Type I extreme value distribution. (The Type I extreme value distribution and density functions are defined, for example, in Eqs. (3.1) and (3.2) Maddala 1983, p. 60.) The resulting choice model is analytically simple but has properties that are undesirable in many applications (for example, the well-known independence-of-irrelevant-alternatives property). Moreover, estimators based on logit models are inconsistent if the distribution of the random components of utility is not Type I extreme value. Manski (1975, 1985) and Cosslett (1983) proposed estimators that do not require a priori knowledge of this distribution. Powell's (1984, 1986) least absolute deviations estimator for censored regression models is another early contribution to econometric research on semiparametric estimation. This estimator was motivated by the observation that estimators of (parametric) Tobit models are inconsistent if the underlying normality assumption is incorrect. Powell's estimator is consistent under very weak distributional assumptions.

Semiparametric estimation has continued to be an active area of econometric research. Semiparametric estimators have been developed for a wide variety of additive, index, partially linear, and hazard models, among others. These estimators all reduce the effective dimension of the estimation problem and overcome the curse of dimensionality by making assumptions that are stronger than those of fully nonparametric estimation but weaker than those of a parametric model. The stronger assumptions also give the models limited extrapolation capabilities. Of course, these benefits come at the price of increased risk of specification error, but the risk is smaller than with simple parametric models. This is because semiparametric models make weaker assumptions than do parametric models, and contain simple parametric models as special cases.

Semiparametric estimation is also an important research field in statistics, and it has led to much interaction between statisticians and econometricians. The early statistics and biostatistics research that is relevant to econometrics was focused on survival (duration) models. Cox's (1972) proportional hazards model and the Buckley and James (1979) estimator for censored regression models are two early examples of this line of research. Somewhat later, Stone (1985) showed that a nonparametric additive model can overcome the curse of dimensionality. Since then, statisticians have contributed actively to research on the same classes of semiparametric models that econometricians have worked on.

Theory-Based Empirical Models

Many econometric models are connected to economic theory only loosely or through essentially arbitrary parametric assumptions about, say, the shapes of utility functions. For example, a logit model of discrete choice assumes that the random components of utility are independently and identically distributed with the Type I extreme value distribution. In addition, it is frequently assumed that the indirect utility function is linear in prices and other characteristics of the alternatives. Because economic theory rarely, if ever, yields a parametric specification of a probability model, it is worth asking whether theory provides useful restrictions on the specification of econometric models, and whether models that are consistent with economic theory can be estimated without making non-theoretical parametric assumptions. The answers to these questions depend on the details of the setting being modelled.

In the case of discrete-choice, random-utility models, the inferential problem is to estimate the distribution of (direct or indirect) utility conditional on observed characteristics of individuals and the alternatives among which they choose. More specifically, in applied research one usually is interested in estimating the systematic

component of utility (that is, the function that gives the mean of utility conditional on the explanatory variables) and the distribution of the random component of utility. Discrete choice is present in a wide range of applications, so it is important to know whether the systematic component of utility and the distribution of the random component can be estimated nonparametrically, thereby avoiding the non-theoretical distributional and functional form assumptions that are required by parametric models. The systematic component and distribution of the random component cannot be estimated unless they are identified. However, economic theory places only weak restrictions on utility functions (for example, shape restrictions such as monotonicity, convexity, and homogeneity), so the classes of conditional mean and utility functions that satisfy the restrictions are large. Indeed, it is not difficult to show that observations of individuals' choices and the values of the explanatory variables, by themselves, do not identify the systematic component of utility and the distribution of the random component without making assumptions that shrink the class of allowed functions.

This issue has been addressed in a series of papers by Matzkin that are summarized in Matzkin (1994). Matzkin gives conditions under which the systematic component of utility and the distribution of the random component are identified without restricting either to a finite-dimensional parametric family. Matzkin also shows how these functions can be estimated consistently when they are identified. Some of the assumptions required for identification may be undesirable in applications. Moreover, Manski (1988) and Horowitz (1998) have given examples in which infinitely many combinations of the systematic component of utility and distribution of the random component are consistent with a binary logit specification of choice probabilities. Thus, discrete-choice, random-utility models can be estimated under assumptions that are considerably weaker than those of, say, logit and probit models, but the systematic component of utility and the distribution of the random component cannot be identified using the restrictions of economic theory alone. It is necessary to make additional assumptions that are not required by economic theory and, because they are required for identification, cannot be tested empirically.

Models of market-entry decisions by oligopolistic firms present identification issues that are closely related to those in discrete-choice, random utility models. Berry and Tamer (2006) explain the identification problems and approaches to resolving them.

The situation is different when the economic setting provides more information about the relation between observables and preferences than is the case in discrete-choice models. This happens in models of certain kinds of auctions, thereby permitting nonparametric estimation of the distribution of values for the auctioned object. An example is a first-price, sealed bid auction within the independent private values paradigm. Here, the problem is to infer the distribution of bidders' values for the auctioned object from observed bids. A game-theory model of bidders' behaviour provides a characterization of the relation between bids and the distribution of private values. Guerre et al. (2000) show that this relation nonparametrically identifies the distribution of values if the analyst observes all bids and certain other mild conditions are satisfied. Guerre et al. (2000) also show how to carry out nonparametric estimation of the value distribution.

Dynamic decision models and equilibrium job-search models are other examples of empirical models that are closely connected to economic theory, though they also rely on non-theoretical parametric assumptions. In a dynamic decision model, an agent makes a certain decision repeatedly over time. For example, an individual may decide each year whether to retire or not. The optimal decision depends on uncertain future events (for example, the state of one's future health) whose probabilities may change over time (for example, the probability of poor health increases as one ages) and depend on the decision. In each period, the decision of an agent who maximizes expected utility is the solution to a stochastic, dynamic programming problem. A large body of research, much of which is reviewed by Rust (1994), shows how to specify and estimate econometric models of the utility

function (or, depending on the application, cost function), probabilities of relevant future events, and the decision process.

An equilibrium search model determines the distributions of job durations and wages endogenously. In such a model, a stochastic process generates wage offers. An unemployed worker accepts an offer if it exceeds his reservation wage. An employed worker accepts an offer if it exceeds his current wage. Employers choose offers to maximize expected profits. Among other things, an equilibrium search model provides an explanation for why seemingly identical workers receive different wages. The theory of equilibrium search models is described in Albrecht and Axell (1984), Mortensen (1990), and Burdett and Mortensen (1998). There is a large body of literature on the estimation of these models. Bowlus et al. (2001) provide a recent example with many references.

The Bootstrap

The exact, finite-sample distributions of econometric estimators and test statistics can rarely be calculated in applications. This is because, except in special cases and under restrictive assumptions (for example, the normal linear model), finite sample distributions depend on the unknown distribution of the population from which the data were sampled. This problem is usually dealt with by making use of large-sample (asymptotic) approximations. A wide variety of econometric estimators and test statistics have distributions that are approximately normal or chi-square when the sample size is large, regardless of the population distribution of the data. The approximation error decreases to zero as the sample size increases. Thus, asymptotic approximations can to be used to obtain confidence intervals for parameters and critical values for tests when the sample size is large.

It has long been known, however, that the asymptotic normal and chi-square approximations can be very inaccurate with the sample sizes encountered in applications. Consequently, there can be large differences between the true and nominal coverage probabilities of confidence intervals and between the true and nominal probabilities with which a test rejects a correct null hypothesis. One approach to dealing with this problem is to use higher-order asymptotic approximations such as Edgeworth or saddlepoint expansions. These received much research attention during 1970s and 1980s, but analytic higher-order expansions are rarely used in applications because of their algebraic complexity.

The bootstrap, which is due to Efron (1979), provides a way to obtain sometimes spectacular improvements in the accuracy of asymptotic approximations while avoiding algebraic complexity. The bootstrap amounts to treating the data as if they were the population. In other words, it creates a pseudo-population whose distribution is the empirical distribution of the data. Under sampling from the pseudo-population, the exact finite sample distribution of any statistic can be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which samples are drawn repeatedly from the empirical distribution of the data. That is, the data are repeatedly sampled randomly with replacement. Since the empirical distribution is close to the population distribution when the sample size is large, the bootstrap consistently estimates the asymptotic distribution of a wide range of important statistics. Thus, the bootstrap provides a way to replace analytic calculations with computation. This is useful when the asymptotic distribution is difficult to work with analytically.

More importantly, the bootstrap provides a low-order Edgeworth approximation to the distribution of a wide variety of asymptotically standard normal and chi-square statistics that are used in applied research. Consequently, the bootstrap provides an approximation to the finite-sample distributions of such statistics that is more accurate than the asymptotic normal or chi-square approximation. The theoretical research leading to this conclusion was carried out by statisticians, but the bootstrap's importance has been recognized in econometrics and there is now an important body of econometric research on the topic. In many settings that are important in applications, the bootstrap essentially eliminates errors in the

coverage probabilities of confidence intervals and the rejection probabilities of tests. Thus, the bootstrap is a very important tool for applied econometricians.

There are, however, situations in which the bootstrap does not estimate a statistic's asymptotic distribution consistently. Manski's (1975, 1985) maximum score estimator of the parameters of a binary response model is an example. All known cases of bootstrap inconsistency can be overcome through the use of subsampling methods. In subsampling, the distribution of a statistic is estimated by carrying out a Monte Carlo simulation in which the subsamples of the data are drawn repeatedly. The subsamples are smaller than the original data-set, and they can be drawn randomly with or without replacement. Subsampling provides estimates of asymptotic distributions that are consistent under very weak assumptions, though it is usually less accurate than the bootstrap when the bootstrap is consistent.

Programme Evaluation and Treatment Effects

Programme evaluation is concerned with estimating the causal effect of a treatment or policy intervention on some population. The problem arises in many disciplines, including biomedical research (for example, the effects of a new medical treatment) and economics (for example, the effects of job training or education on earnings). The most obvious way to learn the effects of treatment on a group of individuals by observing each individual's outcome in both the treated and the untreated states. This is not possible in practice, however, because one virtually always observes any given individual in either the treated state or the untreated state but not both. This does not matter if the individuals who receive treatment are identical to those who do not, but that rarely happens. For example, individuals who choose to take a certain drug or whose physicians prescribe it for them may be sicker than individuals who do not receive the drug. Similarly, people who choose to obtain high levels of education may be different from others in ways that affect future earnings.

This problem has been recognized since at least the time of R.A. Fisher. In principle, it can be overcome by assigning individuals randomly to treatment and control groups. One can then estimate the average effect of treatment by the difference between the average outcomes of treated and untreated individuals. This random assignment procedure has become something of a gold standard in the treatment effects literature. Clinical trials use random assignment, and there have been important economic and social experiments based on this procedure. But there are also serious practical problems. First, random assignment may not be possible. For example, one cannot assign high-school students randomly to receive a university education or not. Second, even if random assignment is possible, post-randomization events may disrupt the effects of randomization. For example, individuals may drop out of the experiment or take treatments other than the one to which they are assigned. Both of these things may happen for reasons that are related to the outcome of interest. For example, very ill members of a control group may figure out that they are not receiving treatment and find a way to obtain the drug being tested. In addition, real-world programmes may not operate the way that experimental ones do, so real-world outcomes may not mimic those found in an experiment, even if nothing has disrupted the randomization.

Much research in econometrics, statistics, and biostatistics has been aimed at developing methods for inferring treatment effects when randomization is not possible or is disrupted by post-randomization events. In econometrics, this research dates back at least to Gronau (1974) and Heckman (1974). The fundamental problem is to identify the effects of treatment or, in less formal terms, to separate the effects of treatment from those of other sources of differences between the treated and untreated groups. Manski (1995), among many others, discusses this problem. Large literatures in statistics, biostatistics, and econometrics are concerned with developing

identifying assumptions that are reasonable in applied settings. However, identifying assumptions are not testable empirically and can be controversial. One widely accepted way of dealing with this problem is to conduct a sensitivity analysis in which the sensitivity of the estimated treatment effect to alternative identifying assumptions is assessed. Another possibility is to forgo controversial identifying assumptions and to find the entire set of outcomes that are consistent with the joint distribution of the observed variables. This approach, which has been pioneered by Manski and several co-investigators, is discussed in Manski (1995, 2003), among other places. Hotz et al. (1997) provide an interesting application of bounding methods to measuring the effects of teenage pregnancy on the labour market outcomes of young women.

Integration and Simulation Methods in Econometrics

The integration problem is endemic in economic modelling, arising whenever economic agents do not observe random variables and the behaviour paradigm is the maximization of expected utility. The econometrician inherits this problem in the expression of the corresponding econometric model, even before taking up inference and estimation. The issue is most familiar in dynamic optimization contexts, where it can be addressed by a variety of methods. Taylor and Uhlig (1990) present a comprehensive review of these methods; for later innovations see Keane and Wolpin (1994), Rust (1997), and Santos and Vigo-Aguiar (1998).

The problem is more pervasive in econometrics than in economic modelling, because it arises, in addition, whenever economic agents observe random variables that the econometrician does not. For example, the economic agent may form expectations conditional on an information set not entirely accessible to the econometrician, such as personal characteristics or confidential information. Another example arises in discrete choice settings, where utilities of alternatives are never observed and the prices of alternatives often are not. In these situations the economic model provides a probability distribution of outcomes conditional on three classes of objects: observed variables, available to the econometrician; latent variables, unobserved by the econometrician; and parameters or functions describing the preferences and decision-making environment of the economic agent. The econometrician typically seeks to learn about the parameters or functions given the observed variables.

There are several ways of dealing with this task. Two approaches that are closely related and widely used in the econometrics literature generate integration problems. The first is to maintain a distribution of the latent variables conditional on observed variables, the parameters in the model, and additional parameters required for completing this distribution. (This is the approach taken in maximum likelihood and Bayesian inference.) Combined with the model, this leads to the joint distribution of outcomes and latent variables conditional on observed variables and parameters. Since the marginal distribution of outcomes is the one relevant for the econometrician in this conditional distribution, there is an integration problem for the latent variables. The second approach is weaker: it restricts to zero the values of certain population moments involving the latent and observable variables. (This is the approach taken in generalized method of moments, which can be implemented with both parametric and nonparametric methods.) These moments depend upon the parameters (which is why the method works) and the econometrician must therefore be able to evaluate the moments for any given set of parameter values. This again requires integration over the latent variables.

Ideally, this integral would be evaluated analytically. Often – indeed, typically – this is not possible. The alternative is to use numerical methods. Some of these are deterministic, but the rapid growth in the solution of these problems since (roughly) 1990 has been driven more by simulation methods employing pseudo-random numbers generated by computer hardware and software. This section reviews the most important

these methods and describes their most significant use in non-Bayesian econometrics, namely, simulated method of moments. In Bayesian econometrics the integration problem is inescapable, the structure of the economic model notwithstanding, because parameters are treated explicitly as unobservable random variables. Consequently simulation methods have been central to Bayesian inference in econometrics.

Deterministic Approximation of Integrals

The evaluation of an integral is a problem as old as the calculus itself. In well-catalogued but limited instances analytical solutions are available: Gradshteyn and Ryzhik (1965) is a useful classic reference. For integration in one dimension there are several methods of deterministic approximation, including Newton-Coates (Press et al. 1986, ch. 4; Davis and Rabinowitz 1984, ch. 2), and Gaussian quadrature (Golub and Welsch 1969; Judd 1998, s. 7.2). Gaussian quadrature approximates a smooth function as the product a polynomial of modest order and a smooth basis function, and then uses iterative refinements to compute the approximation. It is incorporated in most mathematical applications software and is used routinely to approximate integrals in one dimension to many significant figures of accuracy.

Integration in several dimensions by means of deterministic approximation is more difficult. Practical generic adaptations of Gaussian quadrature are limited to situations in which the integrand is approximately the product of functions of single variables (Davis and Rabinowitz 1984, pp. 354–9). Even here the logarithm of computation time is approximately linear in the number of variables, a phenomenon sometimes dubbed 'the curse of dimensionality.' Successful extensions of quadrature beyond dimensions of four or five are rare, and these extensions typically require substantial analytical work before they can be applied successfully.

Low discrepancy methods provide an alternative generic approach to deterministic approximation of integrals in higher dimensions. The approximation is the average value of the integrand computed over a well-chosen sequence of points whose configuration amounts to a sophisticated lattice. Different sequences lead to variants on the approach, the best known being the Halton (1960) sequence and the Hammersley (1960) sequence. Niederreiter (1992) reviews these and other variants.

A key property of any method of integral approximation, deterministic or nondeterministic, is that it should provide as a by-product some indicator of the accuracy of the approximation. Deterministic methods typically provide upper bounds on the approximation error, based on worst-case situations. In many situations the actual error is orders of magnitude less than the upper bound, and as a consequence attaining desired error tolerances may appear to be impractical, whereas in fact these tolerances can easily be attained. Geweke (1996, s. 2.3) provides an example.

Simulation Approximation of Integrals

The structure of integration problems encountered in econometrics makes them often more amenable to attack by simulation methods than by nondeterministic methods. Two characteristics are key. First, integrals in many dimensions are required. In some situations the number is proportional to the size of the sample, and, while the structure of the problem may lead to decomposition in terms of many integrals of smaller dimension, the resulting structure and dimension are still unsuitable for deterministic methods. The second characteristic is that the integration problem usually arises as the need to compute the expected value of a function of a random vector with a given probability distribution P:

$$I = \int_S g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where p is the density corresponding to P, g is the function, $\mathbf{x}$ is the random vector, and I is the number to be approximated. The probability distribution P is then the point of departure for the simulation.

For many distributions there are reliable algorithms, implemented in widely available mathematical applications software, for simulation of random vectors x. This yields a sample $\{g(x^{(m)})\}$

($m = 1, \ldots, M$) whose arithmetic mean provides an approximation of I, and for which a central limit theorem provides an assessment of the accuracy of the approximation in the usual way. (This requires the existence of the first two moments of g, which must be shown analytically.) This approach is most useful when p is simple (so that direct simulation of $\mathbf{x}$ is possible) but the structure of g precludes analytical evaluation of I.

This simple approach does not suffice for the integration problem as it typically arises in econometrics. A leading example is the multinomial probit (MNP) model with J discrete choices. For each individual i the utility of the last choice u_{ij} is normalized to be zero, and the utilities of the first $J - 1$ choices are given by the vector

$$\mathbf{u}_i \sim N(\mathbf{X}_i\beta, \Sigma), \tag{2}$$

where $\mathbf{X}$ is a matrix of characteristics of individual i, including the prices and other properties of the choices presented to that individual, and β and $\mathbf{\Sigma}$ are structural parameters of the model. If the j'th element of $\mathbf{u}_i$ is positive and larger than all the other elements of $\mathbf{u}_i$ the individual makes choice j, and if all elements of $\mathbf{u}$ are negative the individual makes choice J. The probability that individual i makes choice j is the integral of the $(n - 1)$-variate normal distribution (1) taken over the subspace $\{\mathbf{u}_i : u_{ik} \leq u_{ij} \; \forall \; k = 1, \ldots, n\}$. This computation is essential in evaluating the likelihood function, and it has no analytical solution. (For discussion and review, see Sandor and Andras 2004.)

Several generic simulation methods have been used for the problem (1) in econometrics. One of the oldest is acceptance sampling, a simple variant of which is described in von Neumann (1951) and Hammersley and Handscomb (1964). Suppose it is possible to draw from the distribution Q with density q, and the ratio $p(\mathbf{x})/q(\mathbf{x})$ is bounded above by the known constant a. If $\mathbf{x}$ is simulated successively from Q but accepted and taken into the sample with probability $p(\mathbf{x})/[aq(\mathbf{x})]$, then the resulting sample is independently distributed with the identical distribution P. Proofs and further discussion are widely available; for example, Press et al. (1992, s. 7.4), Bratley et al. (1987, s. 5.2.5), and Geweke (2005, s. 4.2.1). The unconditional probability of accepting draws from Q is $1/a$. If a is too large the method is impractical, but when acceptance sampling is practical it provides draws directly from P. This is an important component of many of the algorithms underlying the 'black box' generation of random variables in mathematical applications software.

Alternatively, in the same situation all of the draws from Q are retained and taken into a stratified sample in which the weight $w(\mathbf{x}^{(m)}) = p(\mathbf{x}^{(m)})/q(\mathbf{x}^{(m)})$ is associated with the m'th draw. The approximation of I in (1) is then the weighted average of the terms $g(\mathbf{x}^{(m)})$. This approach dates at least to Hammersley and Handscomb (1964, s. 5.4), and was introduced to econometrics by Kloek and van Dijk (1978). The procedure is more general than acceptance sampling in that a known upper bound of w is not required, but if in fact a is large then the weights will display large variation and the approximation will be poor. This is clear in the central limit theorem for the accuracy of approximation provided in Geweke (1989a), which as a practical matter requires that a finite upper bound on w be established analytically. This is a key limitation of acceptance sampling and importance sampling.

Markov chain Monte Carlo (MCMC) methods provide an entirely different approach to the solution of the integration problem (1). These procedures construct a Markov process of the form

$$\mathbf{x}^{(m)} \sim p\left(\mathbf{x}/\mathbf{x}^{(m-1)}\right) \tag{3}$$

in such a way that

$$M^{-1}\sum_{m=1}^{M} g\left(x^{(m)}\right)$$

converges (almost surely) to I. These methods have a history in mathematical physics dating back to the algorithm of Metropolis et al. (1953). Hastings (1970) focused on statistical problems and extended the method to its present form known as the Hastings–Metropolis (HM) algorithm. HM draws a candidate $\mathbf{x}^*$ from a convenient distribution indexed by $\mathbf{x}^{(m-1)}$. It sets

$\mathbf{x}^{(m)} = \mathbf{x}$ with probability $\alpha(\mathbf{x}^{(m-1)}, \mathbf{x}^{(m)})$ and sets $\mathbf{x}^{(m)} = \mathbf{x}^{(m)-1}$ otherwise, the function α being chosen so that the process (3) defined in this way has the desired convergence property. Chib and Greenberg (1995) provide a detailed introduction to HM and its application in econometrics. Tierney (1994) provides a succinct summary of the relevant continuous state space Markov chain theory bearing on the convergence of MCMC.

A version of the HM algorithm particularly suited to image reconstruction and problems in spatial statistics, known as the Gibbs sampling (GS) algorithm, was introduced by Geman and Geman (1984). This was subsequently shown to have great potential for Bayesian computation by Gelfand and Smith (1990). In GS the vector $\mathbf{x}$ is subdivided into component vectors, $\mathbf{x}' = \left(\mathbf{x}_1', \ldots, \mathbf{x}_B'\right)$, in such a way that simulation from the conditional distribution of each $\mathbf{x}_j$ implied by $p(\mathbf{x})$ in (1) is feasible. This method has proven very advantageous in econometrics generally, and it revolutionized Bayesian approaches in particular beginning about 1990.

By the turn of the century HM and GS algorithms were standard tools for likelihood-based econometrics. Their structure and strategic importance for Bayesian econometrics were conveyed in surveys by Geweke (1999) and Chib (2001), as well as in a number of textbooks, including Koop (2003), Lancaster (2004), Geweke (2005), and Rossi et al. (2005). Central limit theorems can be used to assess the quality of approximations as described in Tierney (1994) and Geweke (2005).

Simulation Methods in Non-Bayesian Econometrics

Generalized method of moments estimation has been a staple of non-Bayesian econometrics since its introduction by Hansen (1982). In an econometric model with $k \times 1$ parameter vector $\boldsymbol{\theta}$ economic theory provides the set of sample moment restrictions

$$\mathbf{h}(\boldsymbol{\theta}) = \int_s \mathbf{g}(\mathbf{x}) p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) d\mathbf{x} = 0, \qquad (4)$$

where $\mathbf{g}(\mathbf{x})$ is a $p \times 1$ vector and $\mathbf{y}$ denotes the data including instrumental variables. An example is the MNP model (2). If the observed choices are coded by the variables $d_{ij} = 1$ if individual i makes choice j and $d_{ij} = 0$ otherwise, then the expected value of d_{ij} is the probability that individual i makes choice j, leading to restrictions of the form (4).

The generalized method of moments estimator minimizes the criterion function $\mathbf{h}(\boldsymbol{\theta})'\mathbf{W}\mathbf{h}(\boldsymbol{\theta})$ given a suitably chosen weighting matrix $\mathbf{W}$. If the requisite integrals can be evaluated analytically, $p \geq k$, and other conditions provided in Hansen (1982) are satisfied, then there is a well-developed asymptotic theory of inference for the parameters that by 1990 was a staple of graduate econometrics textbooks. If for one or more elements of $\mathbf{h}$ the integral cannot be evaluated analytically, then for alternative values of it is often possible to approximate the integral appearing in (4) by simulation. This is the situation in the MNP model.

The substitution of a simulation approximation

$$M^{-1} \sum_{m=1}^{M} \mathbf{g}\left(\mathbf{x}^{(m)}\right)$$

for the integral in (4) defines the method of simulated moments (MSM) introduced by McFadden (1989) and Pakes and Pollard (1989), who were concerned with the MNP model (2) in particular and the estimation of discrete response models using cross-section data in general. Later the method was extended to time series models by Lee and Ingram (1991) and Duffie and Singleton (1993). The asymptotic distribution theory established in this literature requires that the number of simulations M increase at least as rapidly as the square of the number of observations. The practical import of this apparently severe requirement is that applied econometric work must establish that changes in M must have little impact on the results; Geweke et al. (1994, 1997) provide examples for MNP. This literature also shows that in general the impact of using direct simulation, as opposed to analytical evaluation of the integral, is to increase the asymptotic variance of the GMM estimator of $\boldsymbol{\theta}$ by the factor M^{-1}, typically trivial in view of the number of simulations required.

Substantial surveys of the details of MSM and leading applications of the method can be found in Gourieroux and Monfort (1993, 1996), Stern (1997), and Liesenfeld and Breitung (1999).

The simulation approximation, unlike the (unavailable) analytical evaluation of the integral in (4), can lead to a criterion function that is discontinuous in $\boldsymbol{\theta}$. This happens in the MNP model using the obvious simulation scheme in which the choice probabilities are replaced by their proportions in the M simulations, as proposed by Lerman and Manski (1981). The asymptotic theory developed by McFadden (1989) and Pakes and Pollard (1989) copes with this possibility, and led McFadden (1989) to used kernel weighting to smooth the probabilities. The most widely used method for smoothing probabilities in the MNP model is the Geweke–Hajivassiliou–Keane (GHK) simulator of Geweke (1989b), Hajivassiliou et al. (1991), and Keane (1990); a full description is provided in Geweke and Keane (2001), and comparisons of alternative methods are given in Hajivassiliou et al. (1996) and Sandor and Andras (2004).

Maximum likelihood estimation of $\boldsymbol{\theta}$ can lead to first-order conditions of the form (4), and thus becomes a special case of MSM. This context highlights some of the complications introduced by simulation. While the simulation approximation of (1) is unbiased, the corresponding expression enters the log likelihood function and its derivatives nonlinearly. Thus for any finite number of simulations M, the evaluation of the first-order conditions is biased in general. Increasing M at a rate faster than the square of the number of observations eliminates the squared bias relative to the variance of the estimator; Lee (1995) provides further details.

Simulation Methods in Bayesian Econometrics

Bayesian econometrics places a common probability distribution on random variables that can be observed (data) and unobservable parameters and latent variables. Inference proceeds using the distribution of these unobservable entities conditional on the data – the posterior distribution. Results are typically expressed in terms of the expectations of parameters or functions of parameters, expectations taken with respect to the posterior distribution. Thus, whereas integration problems are application-specific in non-Bayesian econometrics, they are endemic in Bayesian econometrics.

The development of modern simulation methods had a correspondingly greater impact in Bayesian than in non-Bayesian econometrics. Since 1990 simulation-based Bayesian methods have become practical in the context of most econometric models. The availability of this tool has been influential in the modelling approach taken in addressing applied econometric problems.

The MNP model (2) illustrates the interaction in latent variable models. Given a sample of n individuals, the $(J-1)\times 1$ latent utility vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are regarded explicitly as $n(J-1)$ unknowns to be inferred along with the unknown parameters β and $\boldsymbol{\Sigma}$. Conditional on these parameters and the data, the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are independently distributed. The distribution of $\mathbf{u}_i$ is (2) truncated to an orthant that depends on the observed choice j: if $j < J$ then $u_{ik} < u_{ij}$ for all $k \neq j$ and $u_{ij} > 0$, whereas for choice J:, $u_{ik} < 0$ for all k. The distribution of each u_{ik}, conditional on all of the other elements of $\mathbf{u}_i$, is truncated univariate normal, and it is relatively straightforward to simulate from this distribution. (Geweke 1991, provides details on sampling from a multivariate normal distribution subject to linear restrictions.) Consequently GS provides a practical algorithm for drawing from the distribution of the latent utility vectors conditional on the parameters.

Conditional on the latent utility vectors – that is, regarding them as observed – the MNP model is a seemingly unrelated regressions model, and the approach taken by Percy (1992) applies. Given conjugate priors the posterior distribution of β, conditional on $\boldsymbol{\Sigma}$ and utilities, is Gaussian, and the conditional distribution of $\boldsymbol{\Sigma}$, conditional on β and utilities, is inverted Wishart. Since GS provides the joint distribution of parameters and latent utilities, the posterior mean of any function of these can be approximated as the sample mean. This approach and the suitability of GS for latent variable models were first recognized by Chib (1992). Similar approaches in other latent variable

models in include McCulloch and Tsay (1994), Chib and Greenberg (1998), McCulloch et al. (2000), and Geweke and Keane (2001).

The Bayesian approach with GS sidesteps the evaluation of the likelihood function and, of any moments in which the approximation is biased given a finite number of simulations, two technical issues that are prominent in MSM. On the other hand, as in all MCMC algorithms, there may be sensitivity to the initial values of parameters and latent variables in the Markov chain, and substantial serial correlation in the chain will reduce the accuracy of the simulation approximation. Geweke (1992, 2005) and Tierney (1994) discuss these issues.

Financial Econometrics

Attempts at testing of the efficient market hypothesis (EMH) provided the impetus for the application of time series econometric methods in finance. The EMH was built on the pioneering work of Bachelier (1900) and evolved in the 1960s from the random walk theory of asset prices advanced by Samuelson (1965). By the early 1970s a consensus had emerged among financial economists suggesting that stock prices could be well approximated by a random walk model and that changes in stock returns were basically unpredictable. Fama (1970) provides an early, definitive statement of this position. He distinguished between different forms of the EMH: the 'weak' form that asserts all price information is fully reflected in asset prices; the 'semi-strong' form that requires asset price changes to fully reflect all publicly available information and not only past prices; and the 'strong' form that postulates that prices fully reflect information even if some investor or group of investors have monopolistic access to some information. Fama regarded the strong form version of the EMH as a benchmark against which the other forms of market efficiencies are to be judged. With respect to the weak form version he concluded that the test results strongly support the hypothesis, and considered the various departures documented as economically unimportant. He reached a similar conclusion with respect to the semi-strong version of the hypothesis. Evidence on the semi-strong form of the EMH was revisited by Fama (1991). By then it was clear that the distinction between the weak and the semi-strong forms of the EMH was redundant. The random walk model could not be maintained either, in view of more recent studies, in particular that of Lo and MacKinlay (1988).

This observation led to a series of empirical studies of stock return predictability over different horizons. It was shown that stock returns can be predicted to some degree by means of interest rates, dividend yields and a variety of macroeconomic variables exhibiting clear business cycle variations. See, for example, Fama and French (1989), Kandel and Stambaugh (1996), and Pesaran and Timmermann (1995) on predictability of equity returns in the United States; and Clare et al. (1994) and Pesaran and Timmermann (2000) on equity return predictability in the UK.

Although it is now generally acknowledged that stock returns could be predictable, there are serious difficulties in interpreting the outcomes of market efficiency tests. Predictability could be due to a number of different factors such as incomplete learning, expectations heterogeneity, time variations in risk premia, transaction costs, or specification searches often carried out in pursuit of predictability. In general, it is not possible to distinguish between the different factors that might lie behind observed predictability of asset returns. As noted by Fama (1991) the test of the EMH involves a joint hypothesis, and can be tested only jointly with an assumed model of market equilibrium. This is not, however, a problem that is unique to financial econometrics; almost all areas of empirical economics are subject to the joint hypotheses problem. The concept of market efficiency is still deemed to be useful as it provides a benchmark and its use in finance has led to significant insights.

Important advances have been made in the development of equilibrium asset pricing models, econometric modelling of asset return volatility (Engle 1982; Bollerslev 1986), analysis of high frequency intraday data, and market microstructures. Some of these developments are reviewed in Campbell et al. (1997), Cochrane (2005),

Shephard (2005), and McAleer and Medeiros (2007). Future advances in financial econometrics are likely to focus on heterogeneity, learning and model uncertainty, real time analysis, and further integration with macroeconometrics. Finance is particularly suited to the application of techniques developed for real time econometrics (Pesaran and Timmermann 2005a).

Appraisals and Future Prospects

Econometrics has come a long way over a relatively short period. Important advances have been made in the compilation of economic data and in the development of concepts, theories and tools for the construction and evaluation of a wide variety of econometric models. Applications of econometric methods can be found in almost every field of economics. Econometric models have been used extensively by government agencies, international organizations and commercial enterprises. Macroeconometric models of differing complexity and size have been constructed for almost every country in the world. In both theory and practice, econometrics has already gone well beyond what its founders envisaged. Time and experience, however, have brought out a number of difficulties that were not apparent at the start.

Econometrics emerged in the 1930s and 1940s in a climate of optimism, in the belief that economic theory could be relied on to identify most, if not all, of the important factors involved in modelling economic reality, and that methods of classical statistical inference could be adapted readily for the purpose of giving empirical content to the received economic theory. This early view of the interaction of theory and measurement in econometrics, however, proved rather illusory. Economic theory is invariably formulated with *ceteris paribus* clauses, and involves unobservable latent variables and general functional forms; it has little to say about adjustment processes, lag lengths and other factors mediating the relationship between the theoretical specification (even if correct) and observables. Even in the choice of variables to be included in econometric relations, the role of economic theory is far more limited than was at first recognized. In a Walrasian general equilibrium model, for example, where everything depends on everything else, there is very little scope for a priori exclusion of variables from equations in an econometric model. There are also institutional features and accounting conventions that have to be allowed for in econometric models but which are either ignored or are only partially dealt with at the theoretical level. All this means that the specification of econometric models inevitably involves important auxiliary assumptions about functional forms, dynamic specifications, latent variables, and so on, with respect to which economic theory is silent or gives only an incomplete guide.

The recognition that economic theory on its own cannot be expected to provide a complete model specification has important consequences for testing and evaluation of economic theories, for forecasting and real time decision making. The incompleteness of economic theories makes the task of testing them a formidable undertaking. In general it will not be possible to say whether the results of the statistical tests have a bearing on the economic theory or the auxiliary assumptions. This ambiguity in testing theories, known as the Duhem–Quine thesis, is not confined to econometrics and arises whenever theories are conjunctions of hypotheses (on this, see for example Cross 1982). The problem is, however, especially serious in econometrics because theory is far less developed in economics than it is in the natural sciences. There are, of course, other difficulties that surround the use of econometric methods for the purpose of testing economic theories. As a rule economic statistics are not the results of designed experiments, but are obtained as by-products of business and government activities often with legal rather than economic considerations in mind. The statistical methods available are generally suitable for large samples while the economic data typically have a rather limited coverage. There are also problems of aggregation over time, commodities and individuals that further complicate the testing of economic theories that are microbased.

Econometric theory and practice seek to provide information required for informed decision-

making in public and private economic policy. This process is limited not only by the adequacy of econometrics but also by the development of economic theory and the adequacy of data and other information. Effective progress, in the future as in the past, will come from simultaneous improvements in econometrics, economic theory and data. Research that specifically addresses the effectiveness of the interface between any two of these three in improving policy – to say nothing of all of them – necessarily transcends traditional sub-disciplinary boundaries within economics. But it is precisely these combinations that hold the greatest promise for the social contribution of academic economics.

Bibliography

Aitken, A.C. 1934–5. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 55: 42–48.

Albrecht, J.W., and B. Axell. 1984. An equilibrium model of search unemployment. *Journal of Political Economy* 92: 824–840.

Allen, R.G.D., and A.L. Bowley. 1935. *Family expenditure*. London: P.S. King.

Almon, S. 1965. The distributed lag between capital appropriations and net expenditures. *Econometrica* 33: 178–196.

Amemiya, T. 1983. Nonlinear regression models. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 1. Amsterdam: North-Holland.

Amemiya, T. 1984. Tobit models: A survey. *Journal of Econometrics* 24: 3–61.

An, S., and F. Schorfheide. 2007. Bayesian analysis of DSGE models. *Econometric Reviews* 26 (2–4): 113–172.

Anderson, T.W., and C. Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Society* 76: 598–606.

Anderson, T.W., and C. Hsiao. 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18: 47–82.

Anderson, T.W., and H. Rubin. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20: 46–63.

Andrews, D.W.K. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61: 821–856.

Andrews, D.W.K., and W. Ploberger. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62: 1383–1414.

Anselin, L. 1988. *Spatial econometrics: Methods and models*. Boston: Kluwer Academic Publishers.

Anselin, L., J. Le Gallo, and H. Jayet. 2007. Spatial panel econometrics. In *The econometrics of panel data: Fundamentals and recent developments in theory and practice*. 3rd ed, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer (forthcoming).

Arellano, M. 2003. *Panel data econometrics*. Oxford: Oxford University Press.

Arellano, M., and S.R. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–297.

Arellano, M., and B. Honoré. 2001. Panel data models: Some recent developments. In *Handbook of econometrics*, ed. J.J. Heckman and E. Leamer, vol. 5. Amsterdam: North-Holland.

Bachelier, L.J.B.A. 1900. *Théorie de la Speculation*. Paris: Gauthier-Villars. Reprinted in *The random character of stock market prices*, ed. P.H. Cootner. Cambridge, MA: MIT Press, 1964.

Bai, J., and S. Ng. 2004. A panic attack on unit roots and cointegration. *Econometrica* 72: 1127–1177.

Bai, J., and P. Perron. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66: 47–78.

Baltagi, B. 2005. *Econometric analysis of panel data*. 2nd ed. New York: Wiley.

Baltagi, B.H., and C. Kao. 2000. Nonstationary panels, cointegration in panels and dynamic panels: A survey. In *Nonstationary panels, panel cointegration, and dynamic panels*, Advances in Econometrics, vol. 15, ed. B. Baltagi. Amsterdam: JAI Press.

Banerjee, A. 1999. Panel data unit roots and cointegration: An overview. *Oxford Bulletin of Economics and Statistics* 61: 607–629.

Basmann, R.L. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25: 77–83.

Bauwens, L., M. Lubrano, and J.F. Richard. 2001. *Bayesian inference in dynamic econometric models*. Oxford: Oxford University Press.

Benini, R. 1907. Sull'uso delle formole empiriche a nell'economia applicata. *Giornale degli economisti,* 2nd series 35: 1053–1063.

Berry, S., and E. Tamer. 2006. Identification in models of oligopoly entry. In *Advances in economics and econometrics: Theory and applications, ninth world congress*, ed. R. Blundell, W.K. Newey, and T. Persson. Cambridge: Cambridge University Press.

Beveridge, S., and C.R. Nelson. 1981. A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary Economics* 7: 151–174.

Binder, M., and M.H. Pesaran. 1995. Multivariate rational expectations models and macroeconometric modelling: A review and some new results. In *Handbook of applied econometrics, vol. 1 – Macroeconomics*, ed. M.H. Pesaran and M.R. Wickens. Oxford: Basil Blackwell.

Binder, M., C. Hsiao, and M.H. Pesaran. 2005. Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory* 21: 795–837.

Bjerkholt, O. 1995. Ragnar Frisch, editor of *Econometrica*. *Econometrica* 63: 755–765.

Blanchard, O.J., and D. Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79: 1146–1164.

Blundell, R.W., M. Browning, and I.A. Crawford. 2003. Nonparametric Engel curves and revealed preference. *Econometrica* 71: 205–240.

Bollerslev, T. 1986. Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* 51: 307–327.

Boskin, M.J. 1974. A conditional logit model of occupational choice. *Journal of Political Economy* 82: 389–398.

Bosq, D. 1996. *Nonparametric statistics for stochastic processes*. New York: Springer.

Bowlus, A.J., N.M. Kiefer, and G.R. Neumann. 2001. Equilibrium search models and the transition from school to work. *International Economic Review* 42: 317–343.

Box, G.E.P., and G.M. Jenkins. 1970. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Bratley, P., B.L. Fox, and L.E. Schrage. 1987. *A guide to simulation*. New York: Springer-Verlag.

Breitung, J., and M.H. Pesaran 2008. Unit roots and cointegration in panels. In *The econometrics of panel data: Fundamentals and recent developments in theory and practice*. 3rd ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Breusch, T.S., and A.R. Pagan. 1980. The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* 47: 239–253.

Brock, W., and S. Durlauf. 2002. A multinomial choice model with neighborhood effects. *American Economic Review* 92: 298–303.

Brock, W., and S. Durlauf. 2006. Macroeconomics and model uncertainty. In *Post-Walrasian macroeconomics: Beyond the dynamic stochastic general equilibrium model*, ed. D. Colander. New York: Cambridge University Press.

Brown, T.M. 1952. Habit persistence and lags in consumer behaviour. *Econometrica* 20: 355–371.

Brown, R.L., J. Durbin, and J.M. Evans. 1975. Techniques for testing the constancy of regression relationships over time (with discussion). *Journal of the Royal Statistical Society, Series B* 37: 149–192.

Broze, L., C. Gouriéroux, and A. Szafarz. 1985. Solutions of dynamic linear rational expectations models. *Econometric Theory* 1: 341–368.

Brundy, J.M., and D.N. Jorgenson. 1971. Efficient estimation of simultaneous equations by instrumental variables. *The Review of Economics and Statistics* 53: 207–224.

Buckley, J., and I. James. 1979. Linear regression with censored data. *Biometrika* 66: 429–436.

Burdett, K., and D.T. Mortensen. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39: 257–273.

Burns, A.F., and W.C. Mitchell. 1947. *Measuring business cycles*. New York: Columbia University Press for the NBER.

Burtless, G. 1995. The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives* 9 (2): 63–84.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Cameron, A.C., and P.K. Trivedi. 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1: 29–53.

Cameron, A.C., and P.K. Trivedi. 1998. *Regression analysis for count data*, Econometric Society Monograph No. 30. Cambridge: Cambridge University Press.

Cameron, A.C., and P.K. Trivedi. 2005. *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Campbell, J.Y., A.W. Lo, and A.C. MacKinlay. 1997. *The econometrics of financial markets*. Princeton: Princeton University Press.

Canova, F., and G. de Nicolo. 2002. Monetary disturbances matter for business fluctuations in the G7. *Journal of Monetary Economics* 49: 1131–1159.

Champernowne, D.G. 1948. Sampling theory applied to autoregressive sequences. *Journal of the Royal Statistical Society, Series B* 10: 204–231.

Champernowne, D.G. 1960. An experimental investigation of the robustness of certain procedures for estimating means and regressions coefficients. *Journal of the Royal Statistical Society* 123: 398–412.

Chib, S. 1992. Bayes inference in the Tobit censored regression model. *Journal of Econometrics* 51: 79–99.

Chib, S. 2001. Markov chain Monte Carlo methods: Computation and inference. In *Handbook of econometrics*, ed. J.J. Heckman and E. Leamer, vol. 5. Amsterdam: North-Holland.

Chib, S., and E. Greenberg. 1995. Understanding the Metropolis–Hastings algorithm. *The American Statistician* 49: 327–335.

Chib, S., and E. Greenberg. 1998. Analysis of multivariate probit models. *Biometrika* 85: 347–361.

Choi, I. 2006. Nonstationary panels. In *Palgrave handbooks of econometrics*, ed. T.C. Mills and K. Patterson, vol. 1. Basingstoke: Palgrave MacMillan.

Chow, G.C. 1960. Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28: 591–605.

Christ, C.F. 1952. *Economic theory and measurement: A twenty-year research report, 1932–52*. Chicago: Cowles Commission for Research in Economics.

Christ, C.F. 1983. The founding of the Econometric Society and Econometrica. *Econometrica* 51: 3–6.

Christiano, L.J., M. Eichenbaum, and C. Evans. 2005. Nominal rigidities and the dynamic effects of a shock

to monetary policy. *Journal of Political Economy* 113: 1–45.

Clare, A.D., S.H. Thomas, and M.R. Wickens. 1994. Is the gilt–equity yield ratio useful for predicting UK stock return? *Economic Journal* 104: 303–315.

Clements, M.P., and D.F. Hendry. 1998. *Forecasting economic time series*. Cambridge: Cambridge University Press.

Clements, M.P., and D.F. Hendry. 1999. *Forecasting non-stationary economic time series*. Cambridge, MA: MIT Press.

Clements, M.P., and D.F. Hendry. 2006. Forecasting with breaks. In *Handbook of economic forecasting*, ed. G. Elliott, C.W.J. Granger, and A. Timmermann, vol. 1. Amsterdam: North-Holland.

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

Cochrane, P., and G.H. Orcutt. 1949. Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association* 44: 32–61.

Conley, T.G., and B. Dupor. 2003. A spatial analysis of sectoral complementarity. *Journal of Political Economy* 111: 311–352.

Conley, T.G., and G. Topa. 2002. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics* 17: 303–327.

Cooper, R.L. 1972. The predictive performance of quarterly econometric models of the United States. In *Econometric models of cyclical behavior*, ed. B.G. Hickman. New York: NBER.

Cosslett, S.R. 1983. Distribution free maximum likelihood estimation of the binary choice model. *Econometrica* 51: 765–782.

Cox, D.R. 1961. Tests of separate families of hypotheses. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1. Berkeley: University of California Press.

Cox, D.R. 1962. Further results of tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B* 24: 406–424.

Cox, D.R. 1972. Regression models and life tables. *Journal of the Royal Statistical Society, Series B* 34: 187–220.

Cross, R. 1982. The Duhem–Quine thesis, Lakatos and the appraisal of theories in macroeconomics. *Economic Journal* 92: 320–340.

Davenant, C. 1698. *Discourses on the public revenues and on the trade of England*, vol. 1. London.

Davidson, R., and J.G. MacKinnon. 2004. *Econometric theory and methods*. Oxford: Oxford University Press.

Davis, P.J., and P. Rabinowitz. 1984. *Methods of numerical integration*. Orlando: Academic.

Deaton, A. 1985. Panel data from time series of cross-sections. *Journal of Econometrics* 30: 109–126.

Del Negro, M., and F. Schorfheide. 2004. Priors from equilibrium models for VAR's. *International Economic Review* 45: 643–673.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters. 2005. *On the fit and forecasting performance of new Keynesian models*, Working Paper No. 491. Frankfurt: European Central Bank.

Dhrymes, P. 1971. A simplified estimator for large-scale econometric models. *Australian Journal of Statistics* 13: 168–175.

Doan, T., R. Litterman, and C.A. Sims. 1984. Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews* 3: 1–100.

Domenich, T., and D. McFadden. 1975. *Urban travel demand: A behavioral analysis*. Amsterdam: North-Holland.

Duffie, D., and K. Singleton. 1993. Simulated moments estimation of Markov models of asset prices. *Econometrica* 61: 929–952.

Duncan, G. 1980. Formulation and statistical analysis of the mixed continuous/discrete variable model in classical production theory. *Econometrica* 48: 839–852.

Durbin, J. 1954. Errors in variables. *Review of the International Statistical Institute* 22: 23–32.

Durbin, J., and G.S. Watson. 1950. Testing for serial correlation in least squares regression I. *Biometrika* 37: 409–428.

Durbin, J., and G.S. Watson. 1951. Testing for serial correlation in least squares regression II. *Biometrika* 38: 159–178.

Durlauf, S.N., P.A. Johnson, and J.R.W. Temple. 2005. Growth econometrics. In *Handbook of economic growth*, ed. P. Aghion and S.N. Durlauf, vol. 1A. Amsterdam: North-Holland.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7: 1–26.

Eisner, R., and R.H. Strotz 1963. Determinants of business investment. In *Impacts of monetary policy.* Englewood Cliffs: Prentice-Hall, for the Commission on Money and Credit.

Elliott, G., C.W.J. Granger, and A. Timmermann. 2006. *Handbook of economic forecasting*. Vol. 1. Amsterdam: North-Holland.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1007.

Engle, R.F. 1984. Wald likelihood ratio and Lagrange multiplier tests in econometrics. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Engle, R.F., and G. Granger. 1987. Cointegration and error-correction: Representation, estimation and testing. *Econometrica* 55: 251–276.

Engle, R.F., D.F. Hendry, and J.-F. Richard. 1983. Exogeneity. *Econometrica* 51: 277–304.

Fachin, S. 2006. Long-run trends in internal migrations in Italy: A study in panel cointegration with dependent units. *Journal of Applied Econometrics* (forthcoming).

Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.

Fama, E.F. 1991. Efficient capital markets: II. *Journal of Finance* 46: 1575–1617.

Fama, E.F., and K.R. French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25: 23–49.

Fan, J., and I. Gijbels. 1996. *Local polynomial modelling and its applications*. London: Chapman & Hall.

Fernandez-Villaverde, J., and J. Rubio-Ramirez. 2005. Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood. *Journal of Applied Econometrics* 20: 891–910.

Fisher, R.A. 1928. *Statistical methods for research workers*. 2nd ed. London: Oliver and Boyd.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan. Reprinted, Philadelphia: Porcupine Press, 1977.

Fisher, I. 1937. Note on a short-cut method for calculating distributed lags. *Bulletin de l'Institut International de Statistique* 29: 323–327.

Fisher, F.M. 1966. *The identification problem in econometrics*. New York: McGraw-Hill.

Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.

Frisch, R. 1933. *Pitfalls in the statistical construction of demand and supply curves*. Leipzig: Hans Buske Verlag.

Frisch, R. 1934. *Statistical confluence analysis by means of complete regression systems*. Oslo: University Institute of Economics.

Frisch, R. 1936. A note on the term 'econometrics'. *Econometrica* 4: 95.

Gali, J. 1992. How well does the IS–LM model fit postwar US data? *Quarterly Journal of Economics* 107: 709–738.

Garratt, A., K. Lee, M.H. Pesaran, and Y. Shin. 2003a. A long run structural macroeconometric model of the UK. *Economic Journal* 113 (487): 412–455.

Garratt, A., K. Lee, M.H. Pesaran, and Y. Shin. 2003b. Forecast uncertainty in macroeconometric modelling: An application to the UK economy. *Journal of the American Statistical Association* 98 (464): 829–838.

Garratt, A., D. Robertson, and S. Wright. 2006a. Permanent vs transitory components and economic fundamentals. *Journal of Applied Econometrics* 21: 521–542.

Garratt, A., K. Lee, M.H. Pesaran, and Y. Shin. 2006b. *Global and national macroeconometric modelling: A long-run structural approach*. Oxford: Oxford University Press.

Geary, R.C. 1949. Studies in relations between economic time series. *Journal of the Royal Statistical Society, Series B* 10: 140–158.

Gelfand, A.E., and A.F.M. Smith. 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85: 398–409.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

Geweke, J. 1989a. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57: 1317–1340.

Geweke, J. 1989b. *Efficient simulation from the multivariate normal distribution subject to linear inequality constraints and the evaluation of constraint probabilities*. Discussion paper, Duke University.

Geweke, J. 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computing science and statistics: Proceedings of the twenty-third symposium on the interface*, ed. E.M. Keramidas. Fairfax: Interface Foundation of North America.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics 4*, ed. J.M. Bernardo et al. Oxford: Clarendon Press.

Geweke, J. 1996. Monte Carlo simulation and numerical integration. In *Handbook of computational economics*, ed. H.M. Amman, D.A. Kendrick, and J. Rust. Amsterdam: North-Holland.

Geweke, J. 1999. Using simulation methods for Bayesian econometric models: Inference, development and communication (with discussion and rejoinder). *Econometric Reviews* 18: 1–126.

Geweke, J. 2005. *Contemporary Bayesian econometrics and statistics*. New York: Wiley.

Geweke, J., and M. Keane. 2001. Computationally intensive methods for integration in econometrics. In *Handbook of econometrics*, ed. J. Heckman and E.E. Leamer, vol. 5. Amsterdam: North-Holland.

Geweke, J., and C. Whiteman. 2006. Bayesian forecasting. In *Handbook of economic forecasting*, ed. G. Elliott, C.W.J. Granger, and A. Timmermann. Amsterdam: North-Holland.

Geweke, J., M. Keane, and D. Runkle. 1994. Alternative computational approaches to statistical inference in the multinomial probit model. *The Review of Economics and Statistics* 76: 609–632.

Geweke, J., M. Keane, and D. Runkle. 1997. Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* 80: 125–165.

Godfrey, L.G. 1988. *Misspecification tests in econometrics: The LM principle and other approaches*. Cambridge: Cambridge University Press.

Godfrey, L.G., and M.R. Wickens. 1982. Tests of misspecification using locally equivalent alternative models. In *Evaluation and reliability of macro-economic models*, ed. G.C. Chow and P. Corsi. New York: Wiley.

Golub, G.H., and J.H. Welsch. 1969. Calculation of Gaussian quadrature rules. *Mathematics of Computation* 23: 221–230.

Gourieroux, C., and J. Jasiak. 2001. *Financial econometrics: Problems, models, and methods*. Oxford: Oxford University Press.

Gourieroux, C., and A. Monfort. 1993. Simulation based inference: A survey with special reference to panel data models. *Journal of Econometrics* 59: 5–33.

Gourieroux, C., and A. Monfort. 1996. *Simulation-based econometric methods*. New York: Oxford University Press.

Gradshteyn, I.S., and I.M. Ryzhik. 1965. *Tables of integrals, series and products*. New York: Academic.

Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37: 424–438.

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 P. Newbold. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2: 111–120.

Granger, C.W.J., and P. Newbold. 1986. *Forecasting economic time series*. 2nd ed. San Diego: Academic.

Granger, C.W.J., and M.H. Pesaran. 2000a. A decision theoretic approach to forecast evaluation. In *Statistics and finance: An interface*, ed. W.S. Chan, W.K. Li, and H. Tong. London: Imperial College Press.

Granger, C.W.J., and M.H. Pesaran. 2000b. Economic and statistical measures of forecast accuracy. *Journal of Forecasting* 19: 537–560.

Greene, W.H. 2003. *Econometric analysis*. 5th ed. New Jersey: Prentice Hall.

Gronau, R. 1974. Wage comparisons – A selectivity bias. *Journal of Political Economy* 82: 1119–1143.

Guerre, E., I. Perrigne, and Q. Vuong. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* 68: 525–574.

Haavelmo, T. 1943. Statistical testing of business cycle theories. *The Review of Economics and Statistics* 25: 13–18.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12 (Supplement): 1–118.

Hajivassiliou, V., D. McFadden, and P. Ruud. 1991. *Simulation of multivariate normal rectangle probabilities. Methods and programs mimeo*. Berkeley: University of California.

Hajivassiliou, V., D. McFadden, and P. Ruud. 1996. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics* 72: 85–134.

Hall, A.R. 2005. *Generalized method of moments*. Oxford: Oxford University Press.

Halton, J.M. 1960. On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik* 2: 84–90.

Hamilton, J.D. 1994. *Time series analysis*. Princeton: Princeton University Press.

Hammersley, J.M. 1960. Monte Carlo methods for solving multivariate problems. *Annals of the New York Academy of Sciences* 86: 844–874.

Hammersley, J.M., and D.C. Handscomb. 1964. *Monte Carlo methods*. London: Methuen.

Hansen, L.P. 1982. Large sample properties of generalized method of moments. *Econometrica* 50: 1029–1054.

Hansen, L.P., and T.J. Sargent. 1980. Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control* 2: 7–46.

Hansen, L.P., and T.J. Sargent. 2007. *Robustness*. Princeton: Princeton University Press.

Härdle, W. 1990. *Applied nonparametric estimation*. Cambridge: Cambridge University Press.

Härdle, W., and O. Linton. 1994. Applied nonparametric methods. In *Handbook of econometrics*, ed. R.F. Engle and D. McFadden, vol. 4. Amsterdam: North-Holland.

Hart, B.S., and J. von Neumann. 1942. Tabulation of the probabilities for the ratio of mean square successive difference to the variance. *Annals of Mathematical Statistics* 13: 207–214.

Harvey, A. 1989. *Forecasting, structural time series models and Kalman Filter*. Cambridge: Cambridge University Press.

Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1272.

Hausman, J.A., and W.K. Newey. 1995. Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica* 63: 1445–1476.

Hausman, J.A., and D.A. Wise, eds. 1985. *Social experimentation*, NBER Conference Report. Chicago: University of Chicago Press.

Hausman, J.A., B.H. Hall, and Z. Griliches. 1984. Econometric models for count data with application to the patents–R&D relationship. *Econometrica* 52: 909–1038.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–694.

Heckman, J.J., and B. Singer. 1984. Econometric duration analysis. *Journal of Econometrics* 24: 63–132.

Heckman, J.J., and A.J. Smith. 1995. Assessing the case for social experimentation. *Journal of Economic Perspectives* 9 (2): 85–110.

Heckman, J.J., and R. Willis. 1977. A beta-logistic model for the analysis of sequential labour force participation by married women. *Journal of Political Economy* 85: 27–58.

Hendricks, K., and R.H. Porter. 1988. An empirical study of an auction with asymmetric information. *American Economic Review* 78: 865–883.

Hendry, D.F., and J.-F. Richard. 1982. On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* 20: 3–33.

Holly, S., M.H. Pesaran, and T. Yamagata. 2006. *A spatio-temporal model of house prices in the US*. Mimeo: University of Cambridge.

Holtz-Eakin, D., W.K. Newey, and H.S. Rosen. 1988. Estimating vector autoregressions with panel data. *Econometrica* 56: 1371–1395.

Honoré, B., and E. Kyriazidou. 2000. Panel data discrete choice models with lagged dependent variables. *Econometrica* 68: 839–874.

Hooker, R.H. 1901. Correlation of the marriage rate with trade. *Journal of the Royal Statistical Society* 44: 485–492.

Horowitz, J.L. 1998. *Semiparametric methods in econometrics*. New York: Springer-Verlag.

Horowitz, J.L. 2001. Semiparametric models. In *International encyclopedia of behavioral and social sciences*, ed. N.J. Smelser and P.B. Baltes. Amsterdam: Elsevier.

Horowitz, J.L., and S. Lee. 2002. Semiparametric methods in applied econometrics: Do the models fit the data? *Statistical Modelling* 2: 3–22.

Horowitz, J.L., and N.E. Savin. 2001. Binary response models: Logits, probits, and semiparametrics. *Journal of Economic Perspectives* 15 (4): 43–56.

Hotz, V.J., C.H. Mullin, and S.G. Sanders. 1997. Bounding causal effects using data from a contaminated natural experiment: Analyzing the effects of teenage childbearing. *Review of Economic Studies* 64: 575–603.

Hsiao, C. 2003. *Analysis of panel data*. 2nd ed. Cambridge: Cambridge University Press.

Hsiao, C., and M.H. Pesaran. 2007. Random coefficient panel data models. In *The econometrics of panel data: Fundamentals and recent developments in theory and practice*. 3rd ed., ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Hsiao, C., M.H. Pesaran, and A.K. Tahmiscioglu. 1999. Bayes estimation of short-run coefficients in dynamic panel data models. In *Analysis of panels and limited dependent variables models*, ed. C. Hsiao et al. Cambridge: Cambridge University press.

Hsiao, C., M.H. Pesaran, and A.K. Tahmiscioglu. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109: 107–150.

Im, K.S., M.H. Pesaran, and Y. Shin. 2003. Testing for unit roots in heterogenous panels. *Journal of Econometrics* 115: 53–74.

Imbs, J., H. Mumtaz, M.O. Ravn, and H. Rey. 2005. PPP strikes back, aggregation and the real exchange rate. *Quarterly Journal of Economics* 120: 1–43.

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12: 231–254. Reprinted in *Long-run economic relationships*, ed. R.F. Engle and C.W.J. Granger. Oxford: Oxford University Press, 1991.

Johansen, S. 1991. Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica* 59: 1551–1580.

Johansen, S. 1995. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford: Oxford University Press.

Jorgenson, D.W. 1966. Rational distributed lag functions. *Econometrica* 34: 135–149.

Judd, K.L. 1998. *Numerical methods in economics*. Cambridge, MA: MIT Press.

Juselius, K. 2006. *The cointegrated VAR model: Econometric methodology and macroeconomic applications*. Oxford: Oxford University Press.

Kagel, J., and A.E. Roth, eds. 1995. *The handbook of experimental economics*. Princeton: Princeton University Press.

Kandel, S., and R.F. Stambaugh. 1996. On the predictability of stock returns: An asset-allocation perspective. *Journal of Finance* 51: 385–424.

Keane, M.P. 1990. *A computationally practical simulation estimator for panel data, with applications to estimating temporal dependence in employment and wages*. Discussion paper, University of Minnesota.

Keane, M., and K.I. Wolpin. 1994. The solution and estimation of discrete choice dynamic programming models by simulation: Monte Carlo evidence. *The Review of Economics and Statistics* 76: 648–672.

Keynes, J.M. 1939. The statistical testing of business cycle theories. *Economic Journal* 49: 558–568.

Kilian, L. 1998. Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics* 80: 218–229.

Kim, K., and A.R. Pagan. 1995. The econometric analysis of calibrated macroeconomic models. In *Handbook of applied econometrics: Macroeconomics*, ed. M.H. Pesaran and M. Wickens. Oxford: Basil Blackwell.

Klein, L.R. 1947. The use of econometric models as a guide to economic policy. *Econometrica* 15: 111–151.

Klein, L.R. 1950. *Economic fluctuations in the United States 1921–1941*, Cowles Commission Monograph No. 11. New York: Wiley.

Kloek, T., and H.K. van Dijk. 1978. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica* 46: 1–20.

Koop, G. 2003. *Bayesian econometrics*. Chichester: Wiley.

Koop, G., and S. Potter. 2004a. *Forecasting and estimating multiple change-point models with an unknown number of change-points*. Mimeo: University of Leicester and Federal Reserve Bank of New York.

Koop, G., and S. Potter. 2004b. *Prior elicitation in multiple change-point models*. Mimeo: University of Leicester and Federal Reserve Bank of New York.

Koop, G., M.H. Pesaran, and S.M. Potter. 1996. Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics* 74: 119–147.

Koopmans, T.C. 1937. *Linear regression analysis of economic time series*. Haarlem: De Erven F. Bohn for the Netherlands Economic Institute.

Koopmans, T.C. 1949. Identification problems in economic model construction. *Econometrica* 17: 125–144.

Koopmans, T.C., H. Rubin, and R.B. Leipnik. 1950. Measuring the equation systems of dynamic economics. In *Statistical inference in dynamic economic models*, Cowles Commission Monograph No. 10, ed. T.C. Koopmans. New York: Wiley.

Koyck, L.M. 1954. *Distributed lags and investment analysis*. Amsterdam: North-Holland.

Laffont, J.-J., H. Ossard, and Q. Vuong. 1995. Econometrics of first-price auctions. *Econometrica* 63: 953–980.

Lancaster, T. 2004. *An introduction to modern Bayesian econometrics*. Malden: Blackwell.

Leamer, E.E. 1978. *Specification searches: Ad Hoc inference with non-experimental data*. New York: Wiley.

Lee, L.F. 1995. Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. *Economic Theory* 11: 437–483.

Lee, B.S., and B. Ingram. 1991. Simulation estimation of time-series models. *Journal of Econometrics* 47: 197–205.

Lenoir, M. 1913. *Etudes sur la formation et le mouvement des prix*. Paris: Giard et Brière.

Lerman, S., and C.S. Manski. 1981. On the use of simulated frequencies to approximate choice probabilities. In *Structural analysis of discrete data with econometric applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Levin, A., C. Lin, and C.J. Chu. 2002. Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics* 108: 1–24.

Liesenfeld, R., and J. Breitung. 1999. Simulation based method of moments. In *Generalized method of moment estimation*, ed. L. Tatyas. Cambridge: Cambridge University Press.

Litterman, R.B. 1985. Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business and Economic Statistics* 4: 25–38.

Liu, T.C. 1960. Underidentification, structural estimation and forecasting. *Econometrica* 28: 855–865.

Lo, A., and C. MacKinlay. 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* 1: 41–66.

Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.

Lucas, R.E. 1973. Some international evidence on output-inflation tradeoffs. *American Economic Review* 63: 326–334.

Lucas, R.E. 1976. Econometric policy evaluation: A critique. In *The Phillips curve and labor markets*, ed. K. Brunner and A.M. Meltzer. Amsterdam: North-Holland.

Lucas, R.E., and T. Sargent 1981. Rational expectations and econometric practice. Introduction to *Rational expectations and econometric practice*. Minneapolis: University of Minnesota Press.

Lűtkepohl, H. 1991. *Introduction to multiple time series analysis*. New York: Springer-Verlag.

Lyttkens, E. 1970. Symmetric and asymmetric estimation methods. In *Interdependent systems*, ed. E. Mosback and H. Wold. Amsterdam: North-Holland.

Maddala, G.S. 1983. *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

Maddala, G.S. 1986. Disequilibrium, self-selection, and switching models. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 3. Amsterdam: North-Holland.

Maddala, G.S. 2001. *Introduction to econometrics*. 3rd ed. New York: Wiley.

Manski, C.F. 1975. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3: 205–228.

Manski, C.F. 1985. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27: 313–334.

Manski, C.F. 1988. Identification of binary response models. *Journal of the American Statistical Association* 83: 729–738.

Manski, C.F. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Manski, C.F. 2003. *Partial identification of probability distributions*. New York: Springer-Verlag.

Manski, C.F., and D. McFadden. 1981. *Structural analysis of discrete data with econometric applications*. Cambridge, MA: MIT Press.

Matzkin, R.L. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: North-Holland.

McAleer, M., and M.C. Medeiros. 2007. Realized volatility: A review. *Econometric Reviews*.

McAleer, M., A.R. Pagan, and P.A. Volker. 1985. What will take the con out of econometrics? *American Economic Review* 75: 293–307.

McCulloch, R.E., and P.E. Rossi. 1994. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* 64: 207–240.

McCulloch, R.E., and R.S. Tsay. 1993. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* 88: 965–978.

McCulloch, R.E., and R.S. Tsay. 1994. Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis* 15: 235–250.

McCulloch, R.E., N.G. Polson, and P.E. Rossi. 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* 99: 173–193.

McFadden, D. 1989. A method of simulated moments for estimation of multinomial probits without numerical integration. *Econometrica* 57: 995–1026.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21: 1087–1092.

Mitchell, W.C. 1928. *Business cycles: The problem in its setting*. New York: NBER.

Moon, R., and B. Perron. 2004. Testing for unit root in panels with dynamic factors. *Journal of Econometrics* 122: 81–126.

Moon, R., and B. Perron. 2007. An empirical analysis of nonstationarity in a panel of interest rates with factors. *Journal of Applied Econometrics* 22: 383–400.

Moore, H.L. 1914. *Economic cycles: Their law and cause*. New York: Macmillan.

Moore, H.L. 1917. *Forecasting the yield and the price of cotton*. New York: Macmillan Press.

Mortensen, D.T. 1990. Equilibrium wage distributions: A synthesis. In *Panel data and labor market studies*, ed. J. Hartog, G. Ridder, and J. Theeuwes. New York: North Holland.

Mundlak, Y. 1961. Empirical production function free of management bias. *Journal of Farm Economics* 43: 44–56.
Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica* 46: 69–85.
Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.
Nadaraya, E.A. 1964. On estimating regression. *Theory of Probability and Its Applications* 10: 141–142.
Nakamura, A., and M. Nakamura. 1981. On the relationships among several specification error tests presented by Durbin, Wu, and Hausman. *Econometrica* 49: 1583–1588.
Nelson, C.R. 1972. The prediction performance of the FRB-MIT-Penn model of the US economy. *American Economic Review* 62: 902–917.
Nerlove, M. 1958a. Adaptive expectations and the cobweb phenomena. *Quarterly Journal of Economics* 72: 227–240.
Nerlove, M. 1958b. *Distributed lags and demand analysis*. Washington, DC: USDA.
Newey, W.K. 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79: 147–168.
Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49: 1399–1416.
Niederreiter, H. 1992. *Random number generation and Quasi-Monte Carol methods*. Philadelphia: SIAM.
Nyblom, J. 1989. Testing for the constancy of parameters over time. *Journal of the American Statistical Association* 84: 223–230.
O'Connell, P. 1998. The overvaluation of purchasing power parity. *Journal of International Economics* 44: 1–19.
Orcutt, G.H. 1948. A study of the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States, 1919–1932 (with discussion). *Journal of the Royal Statistical Society, Series B* 10: 1–53.
Pagan, A.R., and M.H. Pesaran. 2007. On econometric analysis of structural systems with permanent and transitory shocks and exogenous variables. Unpublished manuscript.
Pagan, A., and A. Ullah. 1999. *Nonparametric econometrics*. Cambridge: Cambridge University Press.
Pakes, A., and D. Pollard. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica* 57: 1027–1058.
Pedroni, P. 2001. Purchasing power parity tests in cointegrated panels. *The Review of Economics and Statistics* 83: 727–731.
Pedroni, P. 2004. Panel cointegration: Asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econometric Theory* 20: 597–625.
Peersman, G. 2005. What caused the early millennium slowdown? Evidence based on autoregressions. *Journal of Applied Econometrics* 20: 185–207.
Percy, D.F. 1992. Prediction for seemingly unrelated regressions. *Journal of the Royal Statistical Society, Series B* 54: 243–252.
Pesaran, M.H. 1981. Identification of rational expectations models. *Journal of Econometrics* 16: 375–398.
Pesaran, M.H. 1987a. Econometrics. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.
Pesaran, M.H. 1987b. *The limits to rational expectations*. Oxford: Basil Blackwell.
Pesaran, M.H. 2006. Estimation and inference in large heterogeneous panels with cross section dependence. *Econometrica* 74: 967–1012.
Pesaran, M.H. 2007a. A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Econometrics* 22: 265–312.
Pesaran, M.H. 2007b. A pair-wise approach to testing for output and growth convergence. *Journal of Econometrics* 138: 312–355.
Pesaran, M.H., and A.S. Deaton. 1978. Testing non-nested nonlinear regression models. *Econometrica* 46: 677–694.
Pesaran, M.H., and S. Skouras. 2002. Decision-based methods for forecast evaluation. In *A companion to economic forecasting*, ed. M.P. Cements and D.F. Hendry. Oxford: Blackwell Publishing.
Pesaran, M.H., and R.P. Smith. 1985a. Keynes on econometrics. In *Keynes' economics: Methodological issues*, ed. T. Lawson and M.H. Pesaran. London: Croom Helm.
Pesaran, M.H., and R.P. Smith. 1985b. Evaluation of macroeconometric models. *Economic Modelling* 2: 125–134.
Pesaran, M.H., and R. Smith. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68: 79–113.
Pesaran, M.H., and R.P. Smith. 2006. Macroeconometric modelling with a global perspective. *The Manchester School* 74: 24–49.
Pesaran, M.H., and A. Timmermann. 1995. The robustness and economic significance of predictability of stock returns. *Journal of Finance* 50: 1201–1228.
Pesaran, M.H., and A. Timmermann. 2000. A recursive modelling approach to predicting UK stock returns. *Economic Journal* 110: 159–191.
Pesaran, M.H., and A. Timmermann. 2005a. Real time econometrics. *Econometric Theory* 21: 212–231.
Pesaran, M.H., and A. Timmermann. 2005b. Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics* 129: 183–217.
Pesaran, M.H., and A. Timmermann. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137: 134–161.
Pesaran, M.H., and M. Weeks. 2001. Non-nested hypothesis testing: An overview. In *Companion to theoretical econometrics*, ed. B.H. Baltagi. Oxford: Basil Blackwell.
Pesaran, M.H., T. Schuermann, and S.M. Weiner. 2004. Modelling regional interdependencies using a global error-correcting macroeconometric model (with discussion). *Journal of Business and Economic Statistics* 22 (129–62): 175–181.

Pesaran, M.H., D. Pettenuzzo, and A. Timmermann. 2006. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies* 73: 1057–1084.

Phillips, P.C.B. 1983. Exact small sample theory in the simultaneous equations model. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intrilgator, vol. 1. Amsterdam: North-Holland.

Phillips, P.C.B. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33: 311–340.

Phillips, P.C.B. 1991. Optimal inference in cointegrated systems. *Econometrica* 59: 283–306.

Phillips, P.C.B., and H.R. Moon. 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67: 1057–1111.

Phillips, P.C.B., and Z. Xiao. 1998. A primer on unit root testing. *Journal of Economic Surveys* 12: 423–469.

Powell, J.L. 1984. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25: 303–325.

Powell, J.L. 1986. Censored regression quantiles. *Journal of Econometrics* 32: 143–155.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1986. *Numerical recipes: The art of scientific computing*. 1st ed. Cambridge: Cambridge University Press.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1992. *Numerical recipes: The art of scientific computing*. 2nd ed. Cambridge: Cambridge University Press.

Prothero, D.L., and K.F. Wallis. 1976. Modelling macroeconomic time series. *Journal of the Royal Statistical Society, Series A* 139: 468–486.

Quandt, R.E. 1982. Econometric disequilibrium models. *Econometric Reviews* 1: 1–63.

Ramsey, J.B. 1969. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31: 350–371.

Reiersol, O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9: 1–24.

Reiersol, O. 1945. Confluence analysis by means of instrumental sets of variables. *Arkiv for Mathematik Astronomi och Fysik* 32: 1–119.

Rossi, P.E., G.M. Allenby, and R. McCulloch. 2005. *Bayesian statistics and marketing*. Chichester: Wiley.

Rothenberg, T.J. 1984. Approximating the distributions of econometric estimators and test statistics. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, vol. 2. Amsterdam: North-Holland.

Rust, J. 1994. Structural estimation of Markov decision processes. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: North-Holland.

Rust, J. 1997. Using randomization to break the curse of dimensionality. *Econometrica* 65: 487–516.

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.

Sandor, Z., and P. Andras. 2004. Alternative sampling methods for estimating multivariate normal probabilities. *Journal of Econometrics* 120: 207–234.

Santos, M.S., and J. Vigo-Aguiar. 1998. Analysis of a numerical dynamic programming algorithm applied to economic models. *Econometrica* 66: 409–426.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.

Sargan, J.D. 1964. Wages and prices in the United Kingdom: A study in econometric methodology. In *Econometric analysis for national economic planning*, ed. P.E. Hart, G. Mills, and J.K. Whitaker. London: Butterworths.

Sargent, T.J. 1973. Rational expectations, the real rate of interest and the natural rate of unemployment. *Brookings Papers on Economic Activity* 1973 (2): 429–472.

Sargent, T.J., and N. Wallace. 1975. Rational expectations and the theory of economic policy. *Journal of Monetary Economics* 2: 169–184.

Savin, N.E. 1973. Systems k-class estimators. *Econometrica* 41: 1125–1136.

Schultz, M. 1938. *The theory and measurement of demand*. Chicago: University of Chicago Press.

Schumpeter, J.A. 1954. *History of economic analysis*. London: George Allen & Unwin.

Shephard, N., ed. 2005. *Stochastic volatility: Selected readings*. Oxford: Oxford University Press.

Shiller, R.J. 1973. A distributed lag estimator derived from smoothness priors. *Econometrica* 41: 775–788.

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): 107–164.

Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. In *Problems of economic conditions*, vol. 3. Moscow. English trans. in *Econometrica* 5 (1937): 105–146.

Smets, F., and R. Wouters. 2003. An estimated stochastic dynamic general equilibrium model of the euro area. *Journal of the European Economic Association* 1: 1123–1175.

Smith, V., S. Leybourne, T.-H. Kim, and P. Newbold. 2004. More powerful panel data unit root tests with an application to mean reversion in real exchange rates. *Journal of Applied Econometrics* 19: 147–170.

Solow, R.M. 1960. On a family of lag distributions. *Econometrica* 28: 393–406.

Srivastava, V.K. 1971. Three-stage least-squares and generalized double k-class estimators: A mathematical relationship. *International Economic Review* 12: 312–316.

Stambaugh, R.F. 1999. Predictive regressions. *Journal of Financial Economics* 54: 375–421.

Stern, S. 1997. Simulation-based estimation. *Journal of Economic Literature* 35: 2006–2039.

Stock, J.H. 1994. Unit roots, structural breaks and trends. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Stock, J.H., and M.W. Watson. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* 14: 11–30.

Stock, J.H., J.H. Wright, and M. Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20: 518–529.

Stone, J.R.N. 1945. The analysis of market demand. *Journal of the Royal Statistical Society, Series A* 108: 286–382.

Stone, J.R.N. 1978. *Keynes, political arithmetic and econometrics*. Cambridge: Seventh Keynes Lecture in Economics, British Academy.

Stone, C.J. 1985. Additive regression and other nonparametric models. *Annals of Statistics* 13: 689–705.

Stone, J.R.N. et al. 1954. *Measurement of consumers' expenditures and behavior in the United Kingdom, 1920–38*, 2 vols. London: Cambridge University Press.

Strachan, R.W., and H.K. van Dijk 2006. *Model uncertainty and Bayesian model averaging in vector autoregressive processes*. Discussion Papers in Economics 06/5, Department of Economics, University of Leicester.

Swamy, P.A.V.B. 1970. Efficient inference in a random coefficient regression model. *Econometrica* 38: 311–323.

Taylor, J.B., and H. Uhlig. 1990. Solving nonlinear stochastic growth models: A comparison of alternative solution methods. *Journal of Business and Economic Statistics* 8: 1–18.

Theil, H. 1954. Estimation of parameters of econometric models. *Bulletin of International Statistics Institute* 34: 122–128.

Theil, H. 1958. *Economic forecasts and policy.* 2nd ed. Amsterdam: North-Holland. 1961.

Tierney, L. 1994. Markov chains for exploring posterior distributions with discussion and rejoinder. *Annals of Statistics* 22: 1701–1762.

Tinbergen, J. 1929–30. Bestimmung und Deutung von Angebotskurven: ein Beispiel. *Zeitschrift für Nationalökonomie* 1: 669–679.

Tinbergen, J. 1937. *An econometric approach to business cycle problems*. Paris: Herman & Cie Editeurs.

Tinbergen, J. 1939. *Statistical testing of business cycle theories, vol. 1: A method and its application to investment activity; vol. 2: Business cycles in the United States of America, 1919–1932*. Geneva: League of Nations.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.

Treadway, A.B.. 1971. On the multivariate flexible accelerator. *Econometrica* 39: 845–855.

Trivedi, P.K. 1975. Time series analysis versus structural models: A case study of Canadian manufacturing behaviour. *International Economic Review* 16: 587–608.

Uhlig, H. 2005. What are the effects of monetary policy: Results from an agnostic identification approach. *Journal of Monetary Economics* 52: 381–419.

Verbeek, M. 2007. Pseudo panels and repeated cross-sections. In *The econometrics of panel data: Fundamentals and recent developments in theory and practice*. 3rd ed., ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 12: 367–395.

von Neumann, J. 1942. A further remark on the distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 13: 86–88.

von Neumann, J. 1951. Various techniques used in connection with random digits. *Applied Mathematics Series* 12: 36–38. US National Bureau of Standards.

Wallis, K.F. 1977. Multiple time series analysis and the final form of econometric models. *Econometrica* 45: 1481–1497.

Wallis, K. 1980. Econometric implications of the rational expectations hypothesis. *Econometrica* 48: 49–73.

Watson, G.M. 1964. Smooth regression analysis. *Sankhyâ, Series A* 26: 359–372.

Watson, M.W. 1994. Vector autoregressions and cointegration. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: North-Holland.

Wegge, L.L. 1965. Identifiability criteria for a system of equations as a whole. *Australian Journal of Statistics* 7: 67–77.

West, K.D. 2006. Forecast evaluation. In *Handbook of economic forecasting*, ed. G. Elliott, C. Granger, and A. Timmermann, vol. 1. Amsterdam: North-Holland.

Westerlund, J. 2005. *Panel cointegration tests of the Fisher effect*, Working Papers 2005:10. Department of Economics, Lund University.

White, H. 1981. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76: 419–433.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–26.

Whittle, P. 1963. *Prediction and regulation by linear least-squares methods*. London: English Universities Press.

Wickens, M. 1982. The efficient estimation of econometric models with rational expectations. *Review of Economic Studies* 49: 55–68.

Wickens, M.R., and R. Motto. 2001. Estimating shocks and impulse response functions. *Journal of Applied Econometrics* 16: 371–387.

Wold, H. 1938. *A study in the analysis of stationary time series*. Stockholm: Almqvist and Wiksell.

Wooldridge, J.M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J.M. 2006. *Introductory econometrics: A modern approach*. 3rd ed. Stamford: Thomson-South-Western.

Working, E.J. 1927. What do statistical 'demand curves' show? *Quarterly Journal of Economics* 41: 212–235.
Wright, P.G. 1915. Review of economic cycles by Henry Moore. *Quarterly Journal of Economics* 29: 631–641.
Wright, P.G. 1928. *The tariff on animal and vegetable oils*. London: Macmillan for the Institute of Economics.
Wu, D. 1973. Alternatives tests of independence between stochastic regressor and disturbances. *Econometrica* 41: 733–750.
Yule, G.U. 1895, 1896. On the correlation of total pauperism with proportion of outrelief. *Economic Journal* 5: 603–611; 6: 613–623.
Yule, G.U. 1921. On the time-correlation problem, with special reference to the variate-difference correlation method. *Journal of the Royal Statistical Society* 84: 497–526.
Yule, G.U. 1926. Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* 89: 1–64.
Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57: 348–368.
Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.
Zellner, A. 1984. *Basic issues in econometrics*. Chicago: University of Chicago Press.
Zellner, A. 1985. Bayesian econometrics. *Econometrica* 53: 253–270.
Zellner, A., and F. Palm. 1974. Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* 2: 17–54.
Zellner, A., and H. Theil. 1962. Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica* 30: 54–78.

Economic Anthropology

Timothy Earle

Abstract

Economic anthropology is an empirical science that describes production, exchange and consumption cross-culturally. All societies have economies, but they are variable. Anthropologists evaluate the operations of individual economies and the applicability of Western theories to these cases. Some economic processes work broadly; for example, strategic decision-making, the law of competitive advantage, and calculations of transaction costs help explain many observed patterns. Human economies, however, are often structured as intertwined sectors with distinctive processes. Differences observed in productivity, specialization, institutional structure and social motivations across history and across modern societies are of theoretical significance when constructing the limits of general theory.

Keywords

Behavioural economics; Capitalism; Cognitive ability; Commodity chains; Division of labour; Domestic mode of production; Dual economies; Economic anthropology; Formalism; Gift exchange; Household production; Hunting and gathering economies; Identity; Institutional economics; Markets; Polanyi, K.; Political sector; Postmodernism; Prestige economy; Rationality; Reciprocity; Social networks; Social sector; Specialization; Staple finance and wealth finance; Subsistence sector; Substantivism; Surplus; Technological progress; Transactions costs; Weber, M

JEL Classifications

Z1

Economic anthropology is an empirical science that seeks to describe how production, exchange and consumption operate outside the West (compare Hunt 1997). The second edition (1952) of Herskovits's (1940) text, titled *Economic Anthropology,* labelled this sub-discipline in anthropology. The broader mission of anthropology has been to make sense of the diversity in the human experience, which became apparent to Europeans during progressive stages of exploration, colonialization and globalization. Underlying anthropological research is the premise that human societies have developed parallel institutions of aesthetics, religion, kinship, politics, and of course economics. All societies have economies, and the economic patterns observed in non-Western economies both comfort and confront theories developed by Western scholars.

Common economic processes, such as rational decision-making, law of competitive advantage, and institutional economics help explain many patterns across human economies based on variable conditions of cost, demand and availability. Additionally, however, human economies appear often to be structured quite differently from Western models, and these differences in institutional structure and motivation are of theoretical significance. From the beginning, economic anthropology has contained, and more or less successfully resolved, a tension between the desire to find cross-culturally general theories and to recognize the uniqueness of each individual case. In economic anthropology this tension has been represented in the formalist–substantivist debate.

Few anthropologists identify themselves primarily as economic anthropologists, but study economic matters as part of a broadly integrative approach to human societies. Founded in 1980, the Society of Economic Anthropology is the primary organization for anthropologists with such interests. Members include ethnographers, applied development anthropologists, archaeologists and ethnohistorians, suggesting that economic studies bridge the diversity of the discipline. The society sponsors annual meetings on themes that range across topics including key institutions of labour, property, markets and consumption, and special topics from the gift to slow foods. *Research Series in Economic Anthropology* and *Society for Economic Anthropology Monographs* offer edited volumes on the sub-discipline.

History of Economic Anthropology

From early in the 20th century, anthropologists have questioned whether theories developed to understand Western market economies apply only to those Western societies for which they were generated. To answer this question, anthropologists have described traditional economies, which survived into the 20th century, which existed in the past, and which have been transformed by engagement with the West. Largely empirical, the work is of substantial theoretical significance for understanding economies cross-culturally. Gudeman (1998) has compiled many of the most highly referenced articles.

Economic anthropology's beginning traces to the landmark ethnography *Argonauts of the Western Pacific,* in which Malinowski (1922) described the circulation of shell valuables among the islands of the Kula Ring. Malinowski used the Trobriand Islanders' obsession with certain shell valuables to challenge simplistic notions of 'economic man', and he argued that a non-Western economy could be fundamentally different from modern market economies in values and socialized exchange relationships. Anthropological studies of traditional economies thrived during the first half of the 20th century. As part of British functionalism, Malinowski and his students developed the approach; in French structuralism, Mauss (1925) focused on the gift as a social phenomenon; and, within American anthropology, Herskovits (1940) defined the sub-field. Much of the work was descriptive, emphasizing how traditional people meet basic needs and how the exchange of primitive valuables fashioned and maintained social relationships.

By mid-century, however, studies of traditional economy were increasingly adopting the terms and concepts of Western economic theory. Both Herskovits (1940) and Firth (1939) revised their original books on traditional economies so as to clarify underlying similarities across world economies. They each took concepts, like scarcity and specialization, and generalized them to show that they apply well to societies in which market penetration is not great. They were making the essential point that traditional economies were not simply driven by the food quest. Although most anthropologists took pains to emphasize the differences between traditional economies and market-integrated systems, some seemed to homogenize the human experience, and a sharp reaction followed.

In the tradition of Max Weber, economic historian Karl Polanyi (1944) wrote his famous treatise *The Great Transformation* to argue that the integrating structure of modern markets, for which prices are set by supply and demand, are a very recent creation of industrialism and capitalism. Theories based on scarcity, rationality,

equilibrating price mechanisms operated, he argued, only in the special case of Western capitalism. Modern market conditions should not be taken as inherent in the human experience, but as a recent social artifact malleable in future societies.

Polanyi's impact on economic anthropology was profound and created the debate between substantivists and formalists that raged in the sub-discipline for a generation. *Trade and Markets in the Early Empires* (Polanyi et al. 1957), the seminal edited book, came out of a discussion group which Polanyi led at Columbia University and which included anthropologists who would be influential in the field. Polanyi's chapter 'The Economy as Instituted Process' characterized the substantivist approach. He defined three forms of distribution found in societies with different structured relationships: reciprocity in egalitarian relationships; redistribution in hierarchical relationships; and market exchange in the anonymous relationships of the market. Because economic relations were so deeply embedded in social structure, variation in social organization was thought to explain the differences in the economies. Substantivists recognized that markets were found widely in traditional economies, but argued that those markets were peripheral to most economic activities, which were deeply embedded in social relationships (Bohannan and Dalton 1962). A compendium collected by Dalton (1967) provided empirical cases that illustrate the embedded nature of traditional economies.

In his critique of those using economic theory in non-Western contexts, Polanyi labelled them as 'formalist', meaning that they focused on 'formal' (mathematical) maximizing models to predict how individuals choose among alternative possibilities to allocate limited time, money and other resources. The substantivists, in contrast, focused on how economies were embedded within cultural institutions to meet the material desires that particular culture might have. The debate raged between the two factions through the 1960s and 1970s. Much of the argument became focused on how extensive markets were in traditional societies. In a classic cross-cultural study, Pryor (1977) showed that markets were very broadly distributed, sometimes moving primitive valuable, tools and food. They certainly did not originate with modern capitalism. In his famously acerbic article, Cook (1969) criticized substantivists for being romantic and naive; after all, even if they had useful points to make, the penetration of market economies, he argued, was so pervasive that formalist theories were *now* effectively universal.

Articles representing the two sides were collected in a reader by LeClair and Schneider (1968) that has been used to teach the debate ever since. Articulating the substantivist position, Sahlins (1972) then argued that many concepts of Western economic theory were inapplicable to traditional economy. He discussed the affluence of hunter-gatherers, underproduction in household economies, and the social determinants of reciprocal exchanges. Schneider (1974) countered with the fully articulated formalist position, summarizing how Western economics can be applied cautiously to a wide range of non-Western transactions and decisions, including marriage payments, primitive money, the prestige economy and household production. The debate came to focus on definitions of rationality, scarcity and institutional constraints, but those reading the papers increasingly saw that the participants were talking past each other.

The formalist and substantivist factions represented the inherent tension within anthropology: on the one hand, to seek cross-cultural regularities that reflect shared social process; on the other, to recognize the cultural relativity and uniqueness of each culture. The two sides of the debate fought to exhaustion, as both presented compelling approaches that could be seen as more complementary than alternative. In 1980, Schneider helped organize the Society for Economic Anthropology in order to resolve the debate by bringing the full spectrum of economic anthropologists together. The first meeting, published as *Economic Anthropology*: *Topics and Theories* (Ortiz 1983) gathered an eclectic group of scholars to bridge the theoretical divides within the sub-discipline, with broad interests in marketing, institutions, Marxism, ecology, and economic

development. An edited text, *Economic Anthropology* (Plattner 1989), provided a new generation of students with the breadth of economies and economic conditions that anthropologists were trying to make sense of.

Important to the new harmony has been respect for the different objectives of economic anthropologists, including ethnographic work on traditional economies, applied work on developing economies, and archaeological and historical studies of economies. The field has recognized diversity in both the theoretical and historical nature of human economies. To maintain a proper balance between substantivists and formalists (relativists and universalists) in economic anthropology, the role of archaeological and historical studies has been especially important. As ethnographers increasingly study variants of a single modern system, historical and archaeological studies continue to study the true variation in how human economies are organized and operate. Earle (2002), for example, looks at the alternative means by which political economies have emerged to finance the evolution of chiefdoms and states, showing that the development of market systems is quite rare and specific in that process. Although no careful comparative study exists, the extent of exchange in prehistory appears to have been highly variable.

During the 1980s and 1990s, as economic anthropology matured as a sub-discipline, it became marginalized within anthropology. As in many of the social sciences and humanities, postmodernism became popular, and its anti-materialist, anti-scientific critiques were antithetical to much of what the sub-discipline advocated. As the excesses of postmodernism have receded, however, economic anthropology has regained some of its former popularity, and its potential significance for anthropology and economics seems promising. Perhaps the greatest challenge now is that economics and economic anthropology have remained far apart because of the strongly formal (theoretical) basis of the former and empirical basis of the later. The two approaches would, however, seem complementary.

Economic Anthropology and Its Perspective on World Economies

Economists should consider the empirical value of economic anthropology, and a good place to begin is the compendium *Theory in Economic Anthropology* (Ensminger 2002a). Economic anthropologists are committed to models of reality. The empirical observations and theoretical inferences of anthropology should help recognize the specific frames of applicability for grand theories. In essence, anthropology makes clear that all things are never equal. In this section, I summarize a few conclusions derived from economic anthropology that make a difference to studies of economies. These involve human rationality, consumer behaviour, commodity chains, and the multi-sectored quality of human economies. This list is not meant to be exhaustive, but only to illustrate the importance of cross-cultural evaluations for the models that economists develop. As economics begins to look at such concepts as behavioural economics and personalized networks, the relevance of anthropology's research on these topics becomes particularly significant.

Human decision-making is to a degree rational, and empirical anthropological work significantly improves an understanding of decision-making processes from a cross-cultural and evolutionary perspective. Although rationality underpins much economic theorizing, human cognitive abilities and goals have been under theorized. Recent trends to rectify this within behavioural economics emphasize that individuals do not always act rationally with primary economic objectives and it would appear that economic anthropology could provide valuable cross-cultural validation of these new ideas. Humans prove to be fairly poor decision makers; they appear rather to use simplified proximate measures to estimate such considerations as value and cost (Henrich 2002). Anthropologists have experimented with various economic games given under controlled conditions in non-Western societies, and their results are often counter-intuitive (Ensminger 2002b). In

a sample of societies representing different levels of economic development, for example, as market integration increases cooperation can be shown in such game-playing experiments to become more highly prized.

To understand the evolutionary roots of human rationality, anthropological research has looked at decision-making in small-scale hunting and gathering societies (see for example, Cashdan 1989). As seen by the rapid expansion in brain size deep in history, humans must have been under strong selective pressure for expanded cognitive abilities, and this selective pressure took place when humans were low-density hunter-gatherers. Such hunter-gatherers make daily a wide range of decisions about what foods to eat, where to camp, what groups to join, and the like, and the relative scarcity and abundance of food and their different nutritional qualities appear to be considered. Human cognitive skill determines the ability of hunter-gatherers to adjust rapidly to changing conditions of food availability, to occupy diverse habitats from the Arctic to the tropical forests, and to intensify food procurement as required by population growth. In short, cognitive abilities in the food quest, in movement through the landscape, and in deciding which groups to join must have provided a strong selective advantage that resulted in the moulding of human rationality.

As illustrated by economic anthropology, human decisions often have little direct relationship to economic factors of cost and financial gain. Although of more interest recently to economists, with the notable exception of Thorstein Veblen, economic theory has not attempted systematically to explain how potential consumer outcomes are ranked. Rather, within the West, consumer behaviour has been studied with a rather eclectic and under-theorized set of assumptions. Anthropologists, however, have tried to understand consumption cross-culturally as a social process involving issues of identity and association (Rutz and Orlove 1989). From the anthropological literature, we know how valued objects signify social relationships. The giving and receiving of gifts impart form and meaning to social relationships, and materialize the social distance between actors (Sahlins 1972).

Economic anthropologists frequently study the movements of objects around the globe. These commodity chains describe how goods are produced, distributed and transformed as they move through a sequence of markets (Hansen 2002; Obukhova and Guyer 2002). Commodity chains illustrate how goods, like used clothing, are transformed in value, form and meaning as they pass through a sequence of social worlds and economic sectors. Social considerations of prestige and personal worth are always of great concern in this highly creative process of economic decision-making.

Economic anthropologists have emphasized that economies are multilayered and that the specific character of an economy has historical routes. Although economists often refer to 'dual economy', implying a vestigial survival of traditional practices, they have been reluctant to accept that economies are always multilayered mosaics with spheres of exchange that only partially articulate the different sectors. Economic theory thus radically simplifies reality by focusing on decision-making and outcomes under market conditions, and this simplification makes very different economies appear superficial similar. In the emergence and development of capitalism, since wealth was made in the markets, the primary concern of economists became directed there. As anthropologists seek to understand the different motives and dynamics of economies as articulated in specific social contexts, they have, however, realized that human economies are highly variable, combining subsistence, social, political, and market sectors, each with distinct logics and historical traditions.

The subsistence sector is family-based and involves the daily struggles to meet basic needs. It is universal and represents the economic world of survival in which humans evolved as a species. The primary motivation of humans has probably always been the satisfaction of a family's basic needs. The construction of a general theory of human economies should thus start with how households and communities make a living. Until recently, household requirements were handled largely by family production. Although markets have a long history in human societies, they

were typically quite marginal to subsistence needs. Theorized as the domestic mode of production (DMP; Sahlins 1972), households were oriented to meeting their subsistence needs, and distribution involved sharing between family members with different tasks appropriate to an elementary division of labour by age and gender. In the model, the household is economically self-sufficient, and the economy is not inherently growth-oriented. The amazing conclusion of considerable anthropological research is that the DMP is often at least the model of what the economy should be, and the amounts of goods consumed by households that are produced outside the family have often been but a fraction of the households' overall consumption budget. Prior to the development of full-scale markets, households probably produced 75 per cent or more of everything that they consumed.

The social sector is community-based and involves the lifetime strategies of individuals to define identity and relationships within a broader social group. The social sector is probably universal, finding its roots among early hunter-gatherers and their need to form networks of support, cooperation and exclusion. In cross-cultural perspective, much of the social sector involves reciprocal exchanges within highly social worlds that can be manipulated to emphasize personal prestige. In traditional societies, such competitive exchanges commonly produce social ranking in what has been called a 'prestige economy'. The social sector was elaborated following the Neolithic revolution, as the creation of local corporate groups must have placed a premium on group identity and status. With deep and enduring roots in human history, the social sector would seem to provide a cross-cultural understanding of consumer behaviour as part of processes much broader than capitalism.

Economics now questions assumptions about anonymous markets organized independently of other social institutions. Goods and services are seen as flowing through personalized networks that create the institutions for expanding economic transactions. Greif (2006), for example, argues that the social networks of medieval Europe provided the frameworks for an emergent modern economy. Almost self-evident to anthropologists, such conclusions suggest how economic theory can gain from insights from comparative empirical studies of non-industrial political forms.

Political sectors mobilize and allocate goods to finance regional and interregional institutions of domination and stratification (Earle 2002). Importantly, political economies are not universal. From the fourth millennium bc, the political sector of the economy developed along with chiefdoms and then states. Goods became mobilized as a tax or tribute and then 'redistributed' by dominant political organizations as means to finance their activities. Recent archaeology has studied how political sectors were developed and functioned. An inherent contrast is between staple finance and wealth finance. In staple finance, food goods are mobilized and stored centrally as a means to support craftsmen, warriors and labourers working for the state. Many of these systems, especially in chiefdoms but also some states, functioned with few or no markets. Subsistence and social sectors continued largely unchanged, but new patterns of land ownership and domination required the production of a surplus for ruling institutions. Wealth finance worked similarly, but the local surplus was used to support the production of wealth for tribute payments.

And what about the market sector, so fundamental to most economic theorizing? Archaeological evidence documents that exchange and markets were not universal. From case to case, the amount and types of goods exchanged varied greatly according to specific conditions of availability and production costs and to specific objects of value. Based on ethnographic analogies, until quite recently most of the goods traded were probably handled by down-the-line exchanges between social partners. Goods moving any distance were primarily primitive valuables, items of display and tribute. The extent of exchange in Neolithic and later Bronze Age communities, for example, has been discussed for Europe, where the comparative advantage of one region over another would have been based on the availability of special materials (Sherratt 1997). Subsistence and technological items were rarely exchanged over long distances until the end of the medieval

age. Earlier, some market exchange certainly existed, but their extent and elaboration were apparently quite small.

This empirical record from economic anthropology contests economic theories based on asserted long-term trends in the emergence of marketing. A common assumption among economists from Adam Smith onwards has been that the creation of wealth is an outcome of the development of efficiencies associated with specialization and trade. For example, in his analysis of institutional economics, North (1990) argues that states developed to lower transaction costs between locally specialized but political independent regions. To simplify the logic, technological development and specialization should have created increasing productive economies that, with the emergence of integrating political systems to guarantee the peace of the market, would generate the surplus used to support the growth of civilizations.

The development of markets, however, was quite late and episodic. Following North, economics might suggest that such failure of markets to develop was an outcome of high transaction costs that made exchange unprofitable. Empirically such a conclusion, however, can be shown to be wrong. As political superstructures were developed and imposed broad regional peace that would have radically lowered transactions costs, markets surprisingly did not emerge. The reason appears to be linked to the nature of finance. When finance was based on staples, markets were only rudimentary and peripheral. The complex Hawaiian chiefdoms, for example, conquered and integrated several islands with local specialties in food, stone and other materials, but trade remained very small-scale and local despite the regional peace. Archaeology has documented only minor trade in basaltic adzes and obsidian in Hawaiian prehistory, and these exchanges did not increase with the formation of the large-scale chiefdoms. As a dramatic example, the Inka empire conquered a massive territory that extended 3000 km up the spine of the Andes, imposed an effective regional peace across that territory, and constructed nearly 30,000 km of roads to integrate it. Although these actions would certainly have lowered transaction costs, the regional and distant movements of goods, like metal, ceramics, and foods, remained very limited and completely unchanged from the pre-imperial period (Earle 2002).

Both markets and currencies seem to have expanded in other circumstances where they were linked with wealth finance of states. In the Aztec empire, tribute to the state was in wealth objects like textiles that could be easily transported long distances, centrally stored, and then used as payment to those working for the state. But the use of wealth objects in payment required that the objects be convertible into the staple goods and other consumables desired by state personnel. The Aztec market system provided the mechanism for conversion and was apparently developed by the state (Brumfiel 1980). Afterwards, markets appear to have escaped from state sponsorship and control to take on many of the characterizations commonly associated with market systems.

What are the possibilities for a grand theory of economies? The relatively low status of historical and comparative studies within economics is not promising, but economics would do well to test theories claimed for generally applicability by looking closely at the anthropological literature. To the degree that economic models are used to design economic development in non-Western societies, the general relevance of the economic models must demonstrated. Using a uniform method of analysis, the economist Pryor (2005) has compared industrial economies and traditional (hunter-gatherer and agricultural) economies. His primary conclusions are startling, suggesting the advantages of such comparative analyses. All economies appear to consist of a small number of component parts, probably reflecting the processes and constraints involved in the production and movement of material goods. Economies are thus comparable. Furthermore, the factors that affect such variables as gross productivity or volume of exchange appear not to be determined by social structure but by the particular internal characteristics of the economy. Thus, Polanyi would appear to be wrong; economies are rather independent engines of essential processes. As recent work in economics has relaxed simplifying

assumptions about information, frictionless trade, and anonymity of markets, the potential links between economics and economic anthropology take on reciprocal value.

See Also

- ▶ Behavioural Economics and Game Theory
- ▶ Hunting and Gathering Economies
- ▶ 'Political Economy' and 'Economics'
- ▶ Property Rights
- ▶ Stratification

Bibliography

Bohannan, P., and G. Dalton, eds. 1962. *Markets in Africa*. Evanston: Northwestern University Press.

Brumfiel, E. 1980. Specialization, market exchange, and the Aztec state: A view from Huexotla. *Current Anthropology* 21: 459–478.

Cashdan, E. 1989. Hunters and gatherers: Economic behavior in bands. In *Economic Anthropology*, ed. S. Plattner. Stanford: Stanford University Press.

Cook, S. 1969. The 'anti-market' mentality: A critique of the substantive approach to economic anthropology. *Southwestern Journal of Anthropology* 25: 378–406.

Dalton, G., ed. 1967. *Tribal and peasant economies*. Garden City: Natural History Press.

Earle, T. 2002. *Bronze age economics*. Boulder: Westview.

Ensminger, J., ed. 2002a. *Theory in economic anthropology*. Walnut Creek: AltaMira.

Ensminger, J. 2002b. Experimental economics: A powerful new method for theory testing in anthropology. In Ensminger (2002a).

Firth, R. 1939. *Primitive polynesian economy*, 1965. London: Routledge & Kegan Paul.

Greif, A. 2006. *Institutions and the path to the modern economy: Lessons from medieval trade*. Cambridge: Cambridge University Press.

Gudeman, S. 1998. *Economic anthropology*. Cheltenham: Edward Elgar.

Hansen, K.T. 2002. Commodity chains and the international secondhand clothing trade: *Salaula* and the work of consumption in Zambia. In Ensminger (2002a).

Henrich, J. 2002. Decision-making, cultural transmission and adaptation in economic anthropology. In Ensminger (2002a).

Herskovits, M. 1940. *Economic anthropology*, 1952. New York: Knopf.

Hunt, R. 1997. Economic anthropology. In *The dictionary of anthropology*, ed. T. Barfield. Oxford: Blackwell.

Johnson, A., and T. Earle. 2000. *The evolution of human societies*. 2nd ed. Stanford: Stanford University Press.

LeClair, E.E., and H.K. Schneider, eds. 1968. *Economic anthropology*. New York: Holt, Rinehart.

Malinowski, B. 1922. *Argonauts of the Western Pacific*. London: Routledge.

Mauss, M. 1925. *The gift: Forms and functions of exchange in archaic societies*, 1969. London: Cohen and West.

North, D. 1990. *Institutions, institutional change, and economic performance*. Cambridge: Cambridge University Press.

Obukhova, E., and Guyer, J.I. 2002. Transcending the formal/informal distinction: Commercial relations in Africa and Russia in the post-1989 world. In Ensminger (2002a).

Ortiz, S., ed. 1983. *Economic anthropology: Topics and theories*. Lanham: University Press of America.

Plattner, S., ed. 1989. *Economic anthropology*. Stanford: Stanford University Press.

Polanyi, K. 1944. *The great transformation*. New York: Rinehart.

Polanyi, K., C. Arensberg, and H. Pearson, eds. 1957. *Trade and market in the early empires*. New York: Free Press.

Pryor, F.L. 1977. *The origins of the economy*. New York: Academic.

Pryor, F.L. 2005. *Economic systems of foraging, agricultural, and industrial societies*. Cambridge: Cambridge University Press.

Rutz, H., and B. Orlove, eds. 1989. *The social economy of consumption*. Lanham: University Press of America.

Sahlins, M. 1972. *Stone age economics*. Chicago: Aldine.

Schneider, H.K. 1974. *Economic man: The anthropology of economics*. New York: Free Press.

Sherratt, A. 1997. *Economy and society in prehistoric Europe*. Edinburgh: Edinburgh University Press.

Economic Calculation in Socialist Countries

Michael Ellman

Abstract

In the 1930s, when the classical socialist system emerged, economic decisions were based not on detailed and precise economic methods of calculation but on rough and ready political methods. An important method of economic calculation – particularly in the post-Stalin period – was that of incrementalism. Input norms were a very important method of both inter-industry and consumption planning.

Material balances, and later input–output, were also widely used. Project evaluation, linear programming, comparisons with the West, and economic intuition were other methods used. The influence of methods of economic calculation on economic outcomes should not be exaggerated.

Keywords

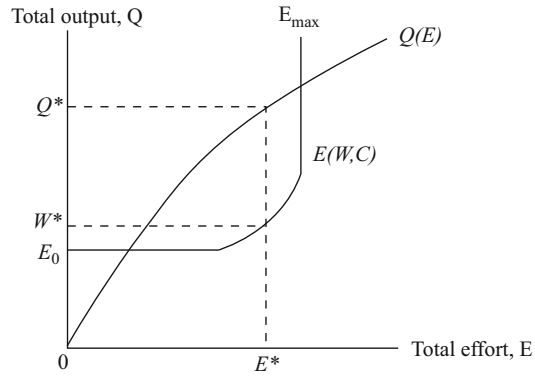
Coefficient of absolute economic effectiveness; Coefficient of relative economic effectiveness; Consumption norms method; Council for mutual economic assistance (Comecon); Economic calculation in social economies; Five year plans (USSR); Gosplan; Gossnab; Incrementalist method of economic calculation; Input norms method of economic calculation; Input–Output method of economic calculation; Kantorovich, L.; Lemeshev, M.; Linear programming; Material balances method; Planned economy; Planning and counter-planning; Stalinism, political economy of; Standard Methodology of economic calculation

JEL Classifications

P21

Economic Calculation and Political Decisions

An important result of the archival revolution of the 1990s (that is, the access to former Soviet archives made possible by the collapse of the USSR) was the additional knowledge it provided about economic decision-making in the USSR in the Stalin period. This made it clear that in the 1930s, when the socialist economic system emerged, economic decisions were based not on detailed and precise economic methods of calculation but on rough and ready political methods. Interesting light has been thrown on the significance of this for macroeconomic, mesoeconomic and microeconomic decision making.



Economic Calculation in Socialist Countries, Fig. 1 Maximizing the investible surplus (*Source*: Adapted from Gregory and Harrison 2005, p. 732)

Macroeconomic policy in the Stalin era aimed to maximize investment subject to the need to provide sufficient consumer goods (mainly food) to maintain labour productivity. The consumer goods were obtained from agriculture by force and allocated by the state in a way which it was hoped would enable investment to be maximized. A schematic representation of short-term macroeconomic calculation under these circumstances is set out in Fig. 1.

Figure 1 shows an output curve OQ which depends on the effort the workers provide, and an effort curve E_oE_{max} which depends on the real wage and the level of coercion. If the state chooses too low a level of wages, output will decline and the intended investment level will be impossible to meet. If wages are set at the fair wage level, output will be maximized but investment less than desired. At the wage level W^*, investment will be maximized. Hence, macroeconomic calculation involved gathering information about worker attitudes (via the state security organizations), allocating the available food to crucial groups of workers, and using coercive or ideological methods to reduce the food–output ratio.

Mesopolicy aimed at developing heavy industry and the defence sector. An important result was what has been termed the 'structural militarization' of the Soviet economy. This resulted from the Soviet view of international relations, the stress on mobilization planning, the lessons

of 1941, and the use by the general staff of absurdly inflated estimates of the mobilization capacity of the USA and other countries. An example is the USSR's capacity at the end of the 1980s to produce about four million tons of aluminium annually. This was greatly in excess of the peacetime economy's need for aluminium. However, in the event of mobilization it would have enabled the country to produce huge numbers of military airplanes. This situation arose as a result of using as a method of economic calculation the attainment of Western levels and of these levels in the military sphere being systematically exaggerated.

On the microeconomic level, Lazarev and Gregory (2003) have studied the allocation of motor vehicles (cars/autos and lorries/trucks) from the central reserve fund in 1932 and 1933. This showed that an economic planning model was unable to explain their allocation (in the regressions the economic variables were insignificant and frequently had the wrong signs). But a political model, in which their allocation was explained as part of a gift-exchange process, explained the data quite well.

Incrementalism

A basic method of economic calculation used in the state socialist countries – particularly in the post-Stalin period – was that of incrementalism, or, as it was known in the USSR, 'planning from the achieved level'. The starting point of all economic plans was the actual or expected outcome of the previous period. The planners adjusted this by reference to anticipated growth rates, current economic policy, shortages and technical progress. For nearly all products, the planned output for next year was the anticipated output for this year plus a few per cent added on. The advantages of incrementalism as a method of economic calculation were its simplicity, realism and compatibility with the functioning of a hierarchical bureaucracy. Its disadvantages were that it provided no method for making technically efficient or consistent decisions, nor did it ensure that the population derived maximum satisfaction from the resources available.

Planning and Counter-Planning

A widely used method of economic calculation was that of planning and counter-planning. If the plan were simply handed down to the enterprises from above, in accordance with the planners' view of national economic requirements but in ignorance of the real possibilities of each enterprise, then it would be unfeasible (if it was too high) or wasteful (if it was too low) or both at the same time (that is, unfeasible for some products and wasteful for others). Conversely, if plans were simply drawn up by each enterprise, they might have failed to use resources in accordance with national economic requirements. The process of planning and counter-planning involved a mutual submission and discussion of planning suggestions, designed to lead to the adoption of a plan which was feasible for the enterprise and ensured that the resources of each enterprise were used in accordance with national requirements.

Unfortunately, the bureaucratic complexity of this procedure militated against both efficiency and consistency.

Input Norms

The main method of economic calculation used to ensure efficiency was that of input norms. An input norm is simply a number assumed to describe an efficient process of transformation of inputs into outputs. For example, suppose that the norm for the utilization of coal in the production of one ton of steel is x tons. Then the efficient production of z tons of steel is assumed to require zx tons of coal.

The method of norms was widely used in Soviet planning, and considerable effort was devoted to updating them. Very detailed norm fixing took place for expenditures of fuel and energy. Much attention was devoted to the development of norms for the expenditure of metal,

cement, and timber in construction. All this work was directed by the department of norms and normatives of Gosplan (the State Planning Commission). Responsibility for elaborating and improving the norms lay with Gosplan's Scientific Research Institute of Planning and Norms.

Nevertheless, the method of norms was incapable of ensuring efficiency. The norms used in planning calculations were simply averages of input requirements, weighted somewhat in favour of efficient producers. Actual technologies showed a wide dispersion in input–output relations. Furthermore, given norms took no account of the possibilities of substitution of inputs for one another in the production process, non-constant returns to scale, and the results of technical progress. Thus in general, the method of norms did not make it possible to calculate efficient input requirements, and plans calculated in this way were always inefficient.

The method of norms was used not only in inter-industry planning but also in consumption planning. In calculating the volume of particular consumer goods and services required, the planners used two main methods. One was forecasts of consumer behaviour, based on extrapolation, expenditure patterns of higher-income groups, income and price elasticities of demand, and consumer behaviour in the more advanced countries. The other method was that of consumption norms. The former method attempted to foresee consumer demand, the latter to shape it.

An example of the method of norms, and its policy implications, is set out in Table 1.

Table 1 makes clear the logic of the Soviet policy in the Brezhnev era (1964–82) of expanding the livestock sector, and also importing fodder and livestock products. Since the consumption of livestock products was below the norm level, the government sought to make possible an increase in their consumption.

The method of consumption norms was an alternative to the price mechanism for the determination of output. It has also been used, however, in Western countries. It is used there in those cases where distribution on the basis of purchasing power has been replaced by distribution on the basis of need. Examples include the provision of housing, hospitals, schools and parks. Calculations of the desirable number of rooms, hospital beds and school places per person are a familiar tool of planning in welfare states.

Economic Calculation in Socialist Countries, Table 1 The soviet diet

	Norm (kgs/head/year)	Per capita consumption in 1976 as % of norm
Bread and bread products	120	128
Potatoes	97	123
Vegetables and melons	37	59
Vegetable oil and margarine	7	85
Meat and meat products	82	68
Fish and fish products	18	101
Milk and milk products	434	78
Eggs	17	72

Sources: Weitzman (1974), Agababyan and Yakovleva (1979, p. 142)

There are two main problems with the norm method of consumption planning. The first is that of substitution between products. Although consumers may well have a medically necessary need for x grams of protein per day, they can obtain these proteins from a wide variety of foods. Second, consumers may choose to spend their money 'irrationally', for example, to buy spirits instead of children's shoes.

Material Balances

A material balance is a balance sheet for a particular commodity showing, on the one hand, the economy's resources and potential output, and, on the other, the economy's need for a particular product. Material (and labour) balances were the main methods used in calculating production and distribution plans for goods, supply plans and labour plans. Soviet planners took great pride in the balance method and considered it one of the greatest achievements of planning theory and practice. Material balances were drawn up for different periods (for example, for annual or five

year periods), by different organizations (for example, Gosplan, Gossnab – the body responsible for allocating supplies of inputs – and the ministries) and at different levels (for example, national and republican). The material balances were also drawn up with different degrees of aggregation. Highly aggregated balances were drawn up for the Five Year Plans, and highly disaggregated balances by the chief administrations of Gossnab for annual supply planning. The aim of the material balance method was to ensure the consistency of the plans.

Normally, at the start of the planning work, the anticipated availability of a commodity was not sufficient to meet anticipated requirements. To balance the two, the planners sought possibilities of economizing on scarce products and substituting for scarce materials; they investigated the possibilities of increasing production or importing raw materials or equipment, or in the last resort they determined the priority needs to be fulfilled by the scarce commodity. Even with great efforts, achieving a balance was difficult. The complexity of an economy in which a great variety of goods are produced by different processes, all of which are subject to continuous technological change, was often too great for anything more than a balance that balanced only on paper. Hence it was normal, during the 'planned' period, for the plan to be altered, often repeatedly, as imbalances came to light. Particularly important problems with the use of material balances were the highly aggregated nature of the balances and their interrelated nature.

Input–Output

A wide variety of input–output tables were regularly constructed in socialist countries. *Ex post* national tables in value terms, planning national tables in value and physical terms, regional tables, and capital stock matrices were widely constructed and used. An interesting and important use concerned variant calculations of the structure of production in medium-term planning.

Because an input–output table can be represented by a simple mathematical model, and because of the assumption of constant coefficients, an input–output table can be utilized for variant calculations.

$$X = (I - A)^{-1} Y$$

On the assumption that A is given, X can be calculated for varying values of Y. Variant calculations of the structure of production were not undertaken with material balances because of their great labour intensity. Variant calculations played a useful role in medium-term planning because they enabled the planners to experiment with a wide range of possibilities. The first major use of variant calculations of the structure of production in Soviet national economic planning was in connection with the 1966–70 Five Year Plan. Gosplan's economic research institute analysed the results of various possible shares of investment in the national income for 1966–70. It became clear that stepping up the share of investment in the national income would increase the rate of growth of the national income, but that this would have very little effect on the rate of growth of consumption (because almost all of the increased output would be producer goods). The results of the calculations are set out in Tables 2 and 3.

Economic Calculation in Socialist Countries, Table 2 Output of steel on various assumptions

	Variants				
	I	II	III	IV	V
Production of steel in 1970 (millions of tonnes)	109	115	121	128	136

Economic Calculation in Socialist Countries, Table 3 Average annual growth rates of selected industries, 1966–1970 (%)

	Variants				
	I	II	III	IV	V
Engineering and metal working	7.1	8.2	9.3	10.4	11.4
Light industry	6.3	6.6	6.8	7.0	7.2
Food industry	7.1	7.3	7.4	7.5	7.6

Source: Ellman (1973, p. 71)

The five variants are for the share of investment in the national income, I being the lowest and V the highest. A sharp increase in the share of investment in the national income in the Five Year Plan 1966–70 would have led to a sharp fall in the share of consumption in the national income, and only a small increase in the rate of growth of consumption (within a Five Year Plan period). What is very sensitive to the share of investment in the national income is the output of the producer goods industries, as Tables 2 and 3 show.

These results are along the lines of what one would expect on the basis of Fel'dman's model, but the input–output technique improves on Fel'dman's model since it enables the effect of different strategies to be seen at the industry level rather than merely in terms of macroeconomic aggregates.

Another example of the use of input–output for economic calculations concerns the statistical data about the relations between industries contained in the national *ex post* tables in value terms. In his controversial 1968 book *Mezhotraslevye svyazi sel'skogo khozyaistva*, M. Lemeshev, then deputy head of the sector for forecasting the development of agriculture of the USSR Gosplan's Economic Research Institute, used the Soviet input–output table for 1959 as the basis for a powerful plea for more industrial inputs to be made available to agriculture.

He began by observing that from the 1959 input–output table it was clear that of the current material inputs into agriculture in that year only 23.4 per cent came from industry, while 54.7 per cent came from agriculture itself (feed, seed and so on). He argued that this was most unsatisfactory. In the section on the relationship between agriculture and engineering Lemeshev argued that the supply to agriculture of agricultural machinery was inadequate, in the section on the relationship between agriculture and the chemical industry he argued that the supply of fertilizers was inadequate, and in the section on agriculture and electricity he argued that the supply of electricity to the villages for both productive and unproductive needs was inadequate. In addition, in the section on the relationship between agriculture and the processing industry he argued that the latter was not helping agriculture as it should do; for example, it was sometimes impossible to accept vegetables (although the consumption of these in the towns was below the norms) because of inadequate processing and distribution facilities. Furthermore, he argued that the supply of concentrated feed was inadequate and the processing of milk wasteful. In view of the inadequate development of the food processing industry, he argued for the development of processing enterprises by the farms themselves.

The chapter on the productive relations between agriculture and the building industry was an extensive critique of the practice of productive, and of housing and communal, building in the villages. Lemeshev argued that the state should take on responsibility for building on the collective farms. The chapter on the relationship between agriculture and transport was critical of the shortage of river freight boats. The chapter on investment argued that investment in agriculture was inadequate, and that in the period 1959–65 there was an unwarranted increase in the proportion of investment in the collective farms which they had to finance themselves. He also argued that a greater proportion of agricultural investment should be financed by bank loans, and that as a criterion of investment efficiency the recoupment period was satisfactory. The concluding chapter was concerned with improving the productive relations between agriculture and the rest of the economy. The author argued for improving central planning by the use of input–output, for replacing procurement plans by free contracts between farms and the procurement organs (if a shortage of a particular product threatened then its price could be raised), and for the elimination of the supply system (that is, the rationing of producer goods) which hindered farms from receiving the goods they wanted and sometimes supplied them with goods that they did not want. Lemeshev also argued for higher pay in agriculture and for the reorganization of the labour process within state and collective farms on the basis of small groups which were paid by results.

This book was a good example of the use of input–output to provide statistical data which could be used, alongside other information, to

provide a description of important economic relations and to support a case for important institutional and policy changes.

Project Evaluation

In the USSR of the 1930s, it was officially considered that there was no problem of project evaluation to which economists could contribute. The sectoral allocation of investment was a matter for the central political leadership to decide. It was they who decided in which sectors and at which locations production should be expanded. These decisions were based on the experience of the more advanced countries, the traditions of the Russian state (for example, stress on railway building) and of the Bolshevik movement (for example, stress on electrification and on the metal-using industries) and on the needs of defence. As far as decisions within sectors were concerned, here the main idea was to fulfil the plan by using the world's most advanced technology.

The practical study of methods for choosing between variants within sectors was begun by engineers in the electricity and railway industries. The problem analysed was that of comparing the cost of alternative ways of meeting particular plan targets. A classic example of the type of problem considered was the choice between producing electricity by a hydro station and by a thermal station.

During Stalin's lifetime, the elaboration by orthodox economists and the adoption by the planners of economic criteria for project evaluation were impossible because they were outside Stalin's conception of the proper role of economists (apologetics). When economists did make a contribution in this area, as was done by Novozhilov, it was ignored. After Stalin's death, however, it became possible for Soviet economists to contribute to the elaboration of methods of economic calculation for use in the decision-making process. An early and important example was in the field of project evaluation. An official method for project evaluation was adopted in 1960, and revised versions in 1964, 1966, 1969 and 1981. In a very abbreviated and summary form, the 1981 version was as follows.

In evaluating investment projects, a wide variety of factors have to be taken into account, for example, the effect of the investment on labour productivity, capital productivity, consumption of current material inputs (such as metals and fuel), costs of production, environmental effects, technical progress, the location of economic activity and so on. Two indices which give useful synthetic information about economic efficiency (but are not necessarily decisive in choosing between investment projects) are the coefficient of absolute economic effectiveness and the coefficient of relative economic effectiveness.

At the national level, the coefficient of absolute effectiveness is defined as the incremental output–capital ratio.

$$E_p = \frac{\Delta Y}{I}$$

where E_p is the coefficient of absolute effectiveness for a particular project, ΔY is the increase in national income generated by the project, and I is the investment cost. The value of E_p calculated in this way for a particular investment has to be compared with E_a, the normative coefficient of absolute effectiveness, which is fixed for each Five Year Plan and varies between sectors. In the 11th Five Year Plan (1981–85) it was 0.16 in industry, 0.07 in agriculture, 0.05 in transport and communications, 0.22 in construction and 0.25 in trade.

$$\text{If } E_p > E_a$$

then the project is considered efficient.

For calculating the criterion of absolute effectiveness at the level of individual industries, net output is used in the numerator instead of national income. At the level of individual enterprises and associations, in particular when a firm's own money or bank loans are the source of finance, profit is used instead of national income.

The coefficient of relative effectiveness is used in the comparison of alternative ways of producing particular products. In the two products case

$$E = \frac{C_1 - C_2}{K_2 - K_1}$$

where E is the coefficient of relative effectiveness, C_i is the current cost of the ith variant, and K_i is the capital cost of the ith variant.

If E > En, where En is the officially established normative coefficient of relative economic efficiency, then the more capital intensive variant is economically justified. In the 11th Five Year Plan, En was in general 0.12, but exceptions were officially permitted in the range 0.08/0.10–0.20/0.25.

In the more than two variants case, they should be compared according to the formula

$$C_i + E_n K_i \rightarrow \text{minimum}$$

that is, choose that variant which minimizes the sum of current and capital costs.

At one time a rationalist misinterpretation of socialist planning was widespread. According to this view, a planned economy was one in which rational decisions were made after a dispassionate analysis by omniscient and all-powerful planners of all the alternative possibilities. In such a system, the adoption of rational criteria for project evaluation would have been of enormous importance. Socialist planning, however, was just one part of the social relations between individuals and groups in the course of which decisions were taken, all of which were imperfect and many of which produced results quite at variance with the intentions of the top economic and political leadership.

A good example of the factors actually influencing investment decisions under state socialism was the commencement of the construction of the Baoshan steel plant near Shanghai. The site was apparently chosen because of the political influence of a high-ranking Shanghai party official. The location decision ignored the fact that, because of the swampy nature of the site, necessitating large expenditures on the foundations, this was in fact the most expensive of the sites considered. Very expensive, dogged with cost overruns, involving major pollution problems, the whole project was kept alive for some time by a powerful steel lobby. In due course, as a result of a national policy reversal in Beijing, the second phase was deferred and those involved publicly criticized. To judge from its initial costs of production, it produced gold rather than steel.

In general, the choice of projects owed more to inter-organization bargaining in an environment characterized by investment hunger than it did to the detached choice of a cost-minimizing variant. The development of new and better criteria for project evaluation turned out to be no guarantee that project evaluation would improve since the criteria were often not in fact used to evaluate projects. Their main function was to provide an acceptable common language in which various bureaucratic agencies conducted their struggles. Agencies adopted projects on normal bureaucratic grounds and then tried to get them adopted by higher agencies, or defended them against attack, by presenting efficiency calculations using the official methodology but relying on carefully selected data.

Linear Programming and Extensions

Linear programming was discovered by the Soviet mathematician Kantorovich in the late 1930s. Its relevance for Soviet planning was widely discussed in the USSR in the 1960s and extensive efforts were made actually to use it in Soviet planning in the 1970s. Three examples of its use follow.

Production Scheduling in the Steel Industry

Linear programming was discovered by Kantorovich in the course of solving the problem, presented to him by the Laboratory of the all-Union Plywood Trust, of allocating productive tasks between machines in such a way as to maximize output given the assortment plan. From a mathematical point of view, the problem of optimal production scheduling for tube mills and rolling mills in the steel industry, which was tackled by Kantorovich in the 1960s, is very similar to the Plywood Trust problem, the difference being its huge dimensions.

The problem arose in the following way. As part of the planning of supply, Soyuzglavmetal (the department of Gossnab concerned with the metal industries), after the quotas had been specified, had to work out production schedules and attachment plans in such a way that all the orders were satisfied and none of the producers received an impossible plan. In the 1960s an extensive research programme was initiated by the department of mathematical economics (which was headed by Academician Kantorovich) of the Institute of Mathematics of the Siberian branch of the Academy of Sciences, to apply optimizing methods to this problem. The chief difficulties were the huge dimensions of the problem and the lack of the necessary data. About 1,000,000 orders, involving 60,000 users, more than 500 producers and tens of thousands of products, were issued each year for rolled metal. Formulated as a linear programming problem it had more than a million unknowns and 30,000 constraints. Collecting the necessary data took about six years. Optimal production scheduling was first applied to the tube mills producing tubes for gas pipelines (these were a scarce commodity in the USSR). In 1970 this made possible an output of tubes 108,000 tons greater than it would otherwise have been, and a substantial reduction in transport costs was also achieved.

The introduction of optimal production scheduling into the work of Soyuzglavmetal was only part of the work initiated in the late 1960s on creating a management information and control system in the steel industry. This was intended to be an integrated computer system which would embrace the determination of requirements, production scheduling, stock control, the distribution of output and accounting. Such systems were widely introduced in Western steel firms in the late 1960s. Work on the introduction of management information and control systems in the Soviet economy was widespread in the 1970s, but by the 1980s there was widespread scepticism in the USSR about their usefulness. This largely resulted from the failure to fulfil the earlier exaggerated hopes about the returns to be obtained from their introduction in the economy.

Industry Investment Plans

In the state socialist countries investment plans were worked out for the country as a whole, and also for industries, ministries, departments, associations, enterprises, republics, economic regions and cities. An important level of investment planning was the industry. Industry investment planning is concerned with such problems as the choice of products, of plants to be expanded, location of new plants, technology to be used, and sources of raw materials.

The main method used in the 1970s and 1980s in the Council for Mutual Economic Assistance (CMEA, known in the West as Comecon) countries for processing the data relating to possible investment plans into actual investment plans was mathematical programming. After extensive experience in this field, in 1977 a Standard Methodology for doing such calculations was adopted by the Presidium of the USSR Academy of Sciences.

The Soviet Standard Methodology presented models for three standard problems. They were: a static multi-product production problem with discrete variables, a multi-product dynamic production problem with discrete variables, and a multi-product static problem of the production-transport type with discrete variables.

The former can be set out as follows:

Let i = 1, ... , n be the finished goods or resources, j = 1,... , m be the production units, r = 1, . . . , Rj be the production technique in a unit, a_{ij}^r be the output of good I = 1 , . . . , n′ or input of resource i = n′ + 1 , . . . , n, using technique r of production in unit j; C_j^r are the costs of production using technique r in unit j; D_i is the given level of output of good i, $I = 1 , . . . , n'$; Pi is the total use of resource i, $i = n' + 1 , . . . , n$ allocated to the industry; Z_j^r is the unknown intensity of use of technique r at unit j.

The problem is to find values of the variables Z_j^r that minimize the objective function

$$\sum_{j=1}^{m} \sum_{j=1}^{R_j} C_j^r Z_j^r \qquad (1)$$

that is, minimize costs of production subject to

$$\sum_{j=1}^{m} \sum_{r=1}^{R_j} a_{ij}^r Z_j^r \geq D_i, \quad i = 1, ..., n' \qquad (2)$$

that is, each output must be produced in at least the required quantities

$$\sum_{j=1}^{m} \sum_{r=1}^{R_j} a_{ij}^r Z_j^r \leq p_i, \quad i = n' + 1, ..., n \qquad (3)$$

that is, the total use of resources cannot exceed the level allocated to the branch

$$\sum_{r=1}^{R_j} Z_j^r \leq 1, j = 1, ..., m \qquad (4)$$

$$Z_j^r = 0 \text{ or } 1, j = 1, ..., m, r = 1, ..., R_j \qquad (5)$$

that is, either a single technique of production for unit j is included in the plan or unit j is not included in the plan.

In order to illustrate the method, an example will be given which is taken from the Hungarian experience of the 1950s in working out an investment plan for the cotton weaving industry for the 1961–65 Five Year Plan. The method of working out the plan can be presented schematically by looking at the decision problems, the constraints, the objective function and the results.

The decision problems to be resolved were:

(a) How should the output of fabrics be increased, by modernizing the existing weaving mills or by building new ones?
(b) For part of the existing machinery, there were three possibilities. It could be operated in its existing form, modernized by way of alterations or supplementary investments, or else scrapped. Which should be chosen?
(c) For the other part of the existing machinery, it could be either retained or scrapped. What should be done?
(d) If new machines are purchased, a choice has to be made between many types. Which types should be chosen, and how many of a particular type should be purchased?

The constraints consisted of the output plan for cloth, the investment fund, the hard currency quota, the building quota and the material balances for various kinds of yarn. The objective function was to meet the given plan at minimum cost.

The results provided answers to all the decision problems. An important feature of the results was the conclusion that it was cheaper to increase production by modernizing and expanding existing mills than by building new ones.

It would clearly be unsatisfactory to optimize the investment plan of each industry taken in isolation. If the calculations show that it is possible to reduce the inputs into a particular industry below those originally envisaged, then it is desirable to reduce planned outputs in other industries, or increase the planned output of the industry in question, or adopt some combination of these strategies. Accordingly, the experiments in working out optimal industry investment plans, begun in Hungary in the 1950s, led to the construction of multi-level plans linking the optimal plans of the separate industries to each other and to the macroeconomic plan variables. Multilevel planning of this type was first developed in Hungary, but subsequently spread to the other CMEA countries. Extensive work on the multi-level optimization of investment planning was undertaken in the USSR in connection with the 1976–90 long-term plan. (The 1976–90 plan, like all previous Soviet attempts to compile a long-term plan, was soon overtaken by events. The plan itself seems never to have been finished and was replaced by ten-year guidelines for 1981–90.)

The Determination of Costs in the Resource Sector

In view of the wide dispersion of production costs in the resource sector, the use of average costs (and of prices based on average costs) in allocation decisions is likely to lead to serious waste. An important outcome of the work of Kantorovich and his school for practical policy was (after a

long lag) official acceptance of this proposition and of linear programming as a way of calculating the relevant marginal costs. For example, in 1979 in the USSR the State Committee for Science and Technology and the State Committee for Prices jointly approved an official method for the economic evaluation of raw material deposits. This was a prescribed method for the economic evaluation of exploration and development of raw material deposits. What was new in principle about this document was that it permitted the output derived from the deposits to be evaluated either in actual (or forecast) wholesale prices or in marginal costs. For the fuel-energy sector, a lot of work was done to calculate actual (and forecast) marginal costs for each fuel at different locations throughout the country and for different periods. These figures were regularly calculated on optimizing models (they were the dual variables to the output maximizing primal) and were widely used in planning practice for many years.

Comparison with the West

An important method of economic calculation in socialist countries was comparison with the West. If a particular product or method of production had already been introduced (or phased out) in the West, this was generally considered a good argument to introduce it (or phase it out) in the socialist countries, subject to national priorities and economic feasibility. Obtaining advanced technology from abroad (by purchase, Lend-Lease, reparations, espionage, direct investment) was an integral part of socialist planning, the importance of the different elements varying over time. Comparisons with the West were particularly important in an economic system which lagged behind the leading countries, lacked institutions which automatically introduced innovations into production (that is, profit-seeking business firms), and found it difficult (because of the ignorance of the planners, stable cost-plus prices and the self-interest of rival bureaucratic agencies) to notice, appraise realistically when noticed, and adopt, innovations.

Economic Calculation and Economic Results

It is important not to exaggerate the influence of methods of economic calculation on the performance of an economy. The performance of an economy is largely determined by external factors (such as the world market), economic policy (for example, the decision to import foreign capital or to declare a moratorium), economic institutions (like collective farms) and the behaviour of the actors within the system (for example, underestimation of investment costs by initiators of investment projects). It is entirely possible for an improvement in the methods of economic calculation to coincide with a worsening of economic performance (as happened in the USSR in the Brezhnev period). Realization of these facts led in the 1970s to a shift from the traditional normative approach (which concentrates on the methods of economic calculation and which regards their improvement as the main key to improved economic performance and the main role of the economist) in the study of planned economies, to the systems and behavioural approaches.

Economic Calculation and Economic Intuition

In view of bounded rationality, and the huge volume, and distorted nature, of the information available to the central leadership, really existing decision-making relied heavily on rules of thumb and the 'feel' for reality of the top decision-makers (sometimes known as 'planning by feel'). This could quickly lead to an equilibrium, but an inefficient one.

See Also

- ▶ Behavioural Public Economics
- ▶ Kantorovich, Leonid Vitalievich (1912–1986)
- ▶ Leontief, Wassily (1906–1999)
- ▶ Soviet Union, Economics in
- ▶ Stalinism, Political Economy of

Bibliography

Agababyan, E., and Ye Yakovleva, eds. 1979. *Problemy raspredeleniya i rost narodnogo blagosostoyaniya*. Moscow: Nauka.

Birman, I. 1978. From the achieved level. *Soviet Studies* 30 (2): 153–172.

Birman, I. 1996. *Otraslevoe optimal'noe. Ch. 9 of Ya –ekonomist*. Novosibirsk: EKOR.

Boltho, A. 1971. *Foreign trade criteria in socialist economies*. Cambridge: Cambridge University Press.

Ellman, M. 1973. *Planning problems in the USSR: The contribution of mathematical economics to their solution 1960–1971*. Cambridge: Cambridge University Press.

Ellman, M. 1983. Changing views on central planning: 1958–1983. *ACES Bulletin [now Comparative Economic Studies]* 25 (1): 11–34.

Gács, J., and M. Lackó. 1973. A study of planning behaviour on the national-economic level. *Economics of Planning* 13: 91–119.

Giffen, J. 1981. The allocation of investment in the Soviet Union. *Soviet Studies* 33 (4): 593–609.

Granick, D. 1990. *Planning as coordination. Ch. 3 of Chinese state enterprises*. Chicago: University of Chicago Press.

Gregory, P., and M. Harrison. 2005. Allocation under dictatorship: Research in Stalin's archives. *Journal of Economic Literature* 43: 721–761.

Kornai, J. 1967. *Mathematical planning of structural decisions*. Amsterdam: North-Holland.

Kornai, J. 1980. *Economics of shortage*, 2 vols. Amsterdam: North-Holland.

Kornai, J. 1992. *Planning and direct bureaucratic control. Ch. 7 of The socialist system*. Oxford: Oxford University Press.

Kueh, Y. 1985. *Economic planning and local mobilization in post-Mao China*. London: Contemporary China Institute.

Kushnirsky, F. 1982. *Soviet economic planning, 1965–1980*. Boulder: Westview, ch. 4.

Lazarev, V., and P. Gregory. 2003. Commissars and cars: A case study in the political economy of dictatorship. *Journal of Comparative Economics* 31: 1–19.

Lemeshev, M. 1968. *Mezhotraslevye svyazi sel'skogo khozyaistva*. Moscow: Ekonomika.

Levine, H. 1959. The centralized planning of supply in Soviet industry. In *Comparisons of the United States and Soviet Economies*. Washington, DC: Joint Economic Committee, US Congress.

Malinovskii, B. 1995. *Istoriya vychislitel'noi tekhniki v litsakh*. Kyiv: KIT/A.S.K.

Matekon. 1978. Standard methodology for calculations to optimize the development and location of production in the long run. 15(1): 75–96.

Qian, Y., G. Roland, and C. Xu. 2000. Coordinating activities under alternative organizational forms. In *Planning, shortage, and transformation*, ed. E. Maskin and A. Simonovits. Cambridge, MA: MIT Press.

Shlykov, V. 2004. The economics of defense in Russia and the legacy of structural militarization. In *The Russian military*, ed. S. Miller and D. Trenin. Cambridge, MA: MIT Press.

Stalin, J. 1952. Concerning the errors of comrade L.D. Yaroshenko. In *Economic problems of socialism in the USSR*. Moscow: Foreign Languages Publishing House.

Tretyakova, A., and I. Birman. 1976. Input–output analysis in the USSR. *Soviet Studies* 28 (2): 157–186.

Weitzman, P. 1974. Soviet long term consumption planning: Distribution according to rational need. *Soviet Studies* 26 (3): 305–321.

World Bank. 1992. *China: Reform and the role of the plan in the 1990s*. Washington, DC: World Bank.

Economic Consequences of Weather, The

Jordan Rappaport

Abstract

Households in the United States and a number of other wealthy nations have been migrating to places with nice weather. This likely reflects an increase in the relative valuation of the weather's direct contribution to household utility. Several different amenity explanations are discussed that can account for the increased valuation and ongoing move.

Keywords

Compensating differentials; Consumption amenities; Local growth; Migration; Weather

JEL Classifications

R11; R12; R13; R23

Introduction

> A cloudy day or a little sunshine have as great an influence on many constitutions as the most recent blessings or misfortunes.
> Joseph Addison (1672–1719), English essayist, poet and politician

> Don't knock the weather. If it didn't change once in a while, nine out of ten people couldn't start a conversation.
> Kin Hubbard (1868–1930) American cartoonist, humorist and journalist

It is hard to find a research subject more important than the weather. From ice ages to epic floods to endless droughts to malarial heat to the present warming of the earth, human welfare has always depended closely on it. Less awe-inspiring but also important is that weather is a direct source of significant consumption. Nice weather underpins the enjoyment of most outdoor activities from picnics to sports games to beach days, to an infinite set of other possibilities. The discussion that follows will focus primarily on this latter, consumption dimension of weather. While such a focus may seem shallow in the face of the significant challenges weather poses to humanity, those challenges do not negate the fact that normal weather variations – the sorts that have been experienced year after year by current and recent generations – continue to be a large source of consumption benefits.

The discussion below will argue that rising incomes in the United States and other developed nations have increased households' willingness to pay to live in a place with nice weather. As a result there has been a shift in population towards such places. Before we consider this consumption dimension of weather, however, a brief discussion of the weather's day-to-day contribution to production is warranted.

Weather as a Production Amenity

Agriculture is the industry that most obviously depends on weather as a productive input. This dependence is multidimensional in the sense that temperature, humidity, cloud cover and rainfall – each over the entire growing season – all matter. A large enough deviation by just one of these can be sufficient to seriously impair yields. To be sure, advancing agricultural science has allowed crops to thrive in a wider range of weather conditions. But even loosened, the constraints imposed by weather remain significant.

Of course, different agricultural goods thrive in different weather. But abstracting from heterogeneity, it is easy to see that farmland in places with weather most conducive to growing will be valued especially highly. Farmers, assumed to be mobile across locations, will bid up the price of productive farmland until the weather's expected contribution to profits becomes fully capitalized into land values. The higher productivity of farms in ideal-weather locations simultaneously makes it possible to pay workers there relatively high wages while still attaining the profits that could be made elsewhere. Note that the farm workers in such high productivity locations are not necessarily any better off than mobile farm workers elsewhere. General equilibrium considerations imply that their higher wages will be offset by higher prices for non-traded goods such as housing.

As with agriculture, the weather serves as a productive input into numerous industrial processes. Gunpowder, macaroni, tobacco, gum and chocolate are among the many products whose production requires constant, low humidity. Inside weather conditions are thus an extremely important productive input. Of course, in present-day developed countries, inside and outside weather are typically disconnected. But prior to air conditioning and central heating, inside weather depended closely on outside weather. Thus Oi (1997) argues that the spread of workplace air conditioning underpinned the rise of manufacturing in the south of the United States.

Nice weather also turns out to be empirically correlated with very-short-term stock market returns (Saunders 1993; Hirshleifer and Shumway 2003). Specifically, daily measures of sunshine in cities that host major stock exchanges are positively correlated with daily returns on those exchanges. In this case, weather's contribution is potentially productive only for day traders with very low transaction costs. The hypothesized mechanism is that nice weather uplifts traders' mood and optimism. Such a mechanism has much more the flavour of a consumption amenity than a productive one.

Weather as a Consumption Amenity

Just as weather's contribution to production puts upward pressure on the price of land and of

housing, so too does its direct contribution to household utility. But as a consumption amenity, weather puts downward pressure on wages rather than upward pressure.

The expected correlations from weather's role as a production amenity and as a consumption amenity derive from the compensating differential framework (Rosen 1979; Roback 1982). An economy is assumed to be made up of a number of geographically distinct labour markets where households live and work and firms produce. The labour market locations may differ from one another with respect to numerous exogenous production and consumption amenities such as proximity to navigable water, access to natural resources, low risk of natural disasters, and – in a multidimensional sense – the weather. Production and consumption amenities may also be endogenous, for instance if increasing returns to scale lower input costs or expand the variety of consumer goods. The assumed high mobility of firms implies that they must be at least as profitable in their present location as they would be anywhere else. The assumed high mobility of households implies that they must derive at least as much utility in their present location as they would anywhere else.

The key to the compensating differential framework is that prices – in particular for land, labour and housing services – adjust to equate profits and utility across the numerous locations. In locations with high production amenities, firms are willing to pay higher prices for inputs, including for labour. These higher input prices are required to lower what would otherwise be higher profits than could be achieved from locating elsewhere. Similarly, in locations with high consumption amenities, households are willing to accept lower wages and pay a higher price for housing services. Such households thus trade off lower tangible consumption of market goods for higher intangible consumption of amenities.

The empirical implementation of this model typically focuses exclusively on households and consumption amenities rather than firms and production amenities. The reason is the difficulty of observing the full range of firm input prices. Notable exceptions include Gabriel and Rosenthal (2004) and Chen and Rosenthal (2008), which treat housing service prices as a proxy for non-labour input prices.

For households, the most common empirical methodology is to separately regress micro data of household income and a proxy for housing service price on respective vectors of attributes meant to control for differences in human capital and differences in the quantity and quality of housing services. The residuals from these regressions can then be regressed on location-specific attributes, including weather. Summing the extra annual housing service cost implied by a coefficient on a locational attribute in the housing regression with the lost income implied by the coefficient on the same locational attribute in the income regression gives the marginal consumption that a household forgoes to obtain a small increase in that local attribute.

Estimated compensating differentials for weather attributes from implementing this methodology tend to be extremely large. For example, the valuation per representative household for one extra sunny day over the course of a year is somewhere from US$21 (at 2005 prices) to $36. The midpoint of this estimated range implies an aggregate valuation of $57 million per year for a metropolitan area with a population of 2 million. Over 30 years using a three per cent discount rate, the implied net present value is $560 million. Whether households really require such a huge transfer to accept just a single extra cloudy day per year seems questionable. Other estimated weather valuations include one less rainy day over the course of a year, $36 per household; one less inch of precipitation, $-63 to $37 per household; and one inch less snow per year, $33 per household (Blomquist et al. 1988; Gyourko and Tracy 1991; Stover and Leven 1992).

Heterogeneity of household preferences suggests that these estimates may understate the consumption benefits from weather. With heterogeneity, it is no longer necessary that all households be indifferent about where to live. The distribution of wages and house prices across locations that clears the labour, traded goods, and housing markets will be driven in large part by 'marginal' households, who tend to value

consumption amenities by less than average. 'Inframarginal' households, in contrast, tend to value at least some consumption amenities highly. In order to live in a location where such amenities are abundant, inframarginal households are willing to accept a lower wage and pay a higher housing-service price than is actually required. Hence they enjoy a surplus that is missed by the compensating valuations above.

An even bigger empirical challenge to valuing weather and other consumption amenities is the difficulty of controlling for individual-specific and house-specific characteristics. A low wage may represent compensation for amenities, but it also may represent low human capital. A high expenditure on housing may compensate for high amenities, but it also may reflect a high quality and quantity of housing services being purchased. The characteristics typically used as controls when estimating the wage compensation include age, experience, education, sex, industry and occupation. For estimating the house price compensation, typical controls include rooms, bedrooms, units in structure, and appliances. These sets of attributes miss substantial sources of individual and housing-unit variation. Probably most important for present purposes is the difficulty of distinguishing between high amenities and low human capital. The sorting of human capital across metro areas suggests that unobserved human capital characteristics may be correlated with the weather. The consequences of not sufficiently controlling for individual and housing service characteristics are evident in quality-of-life rankings of metro areas based on compensating differentials, which tend to contrast sharply with subjective rankings (Rappaport 2008).

A complementary 'quantity' approach to the compensating differential literature's 'price' approach explicitly models population, capital inputs, land and housing supply. As is intuitive, high levels of consumption amenities attract households to a location, resulting in higher population and population density. (Henceforth, I shall make no distinction between the level of population and its density.) The higher population in turn supports the higher housing prices and lower wages of the compensating equilibrium (Haurin 1980; Rappaport 2008).

The seemingly obvious empirical implication of the quantity approach is to regress a cross-section of local population on exogenous local attributes such as the weather to infer whether such attributes are an amenity (with respect to either production or consumption). However, the extremely high persistence of local population implies that the correlation of population with an attribute might reflect an amenity contribution in the distant past that no longer exists. Instead, a crosssection of population growth rates can be regressed on the exogenous attributes. The resulting coefficients can be interpreted as reflecting the accumulation of past changes of the attributes' amenity contributions (Mueser and Graves 1995; Rappaport 2007). In other words, a positive partial correlation between population growth and a particular attribute suggests that the attribute's amenity contribution increased – becoming either more positive or less negative – in the intermediate past. The high persistence of population growth in the United States suggests that the 'intermediate past' probably reaches back at least several decades (Greenwood et al. 1991; Rappaport 2004; Glaeser and Gyourko 2005).

Empirically implementing the quantity approach establishes that population growth in the United States has been highly correlated with nice weather. Growth has been fastest where winters and summers are mild and the number of rainy days is moderate. The quantitatively strongest relationship, robust to numerous controls, is a positive quadratic correlation of growth with winter temperature. For the period 1970 to 2000, increasing January temperature from one standard deviation below its sample mean to one standard deviation above its sample mean (from 29 °F to 54 °F) is associated with faster growth of 1.3 per cent per year for US counties (Rappaport 2007). Miami's temperature in January implies expected annual growth that is 3.4 per cent faster than that of US counties with mean January temperature. For comparison, the mean population growth rate of counties over this period was 0.9 per cent per year.

Population growth is negatively correlated with summer temperature and humidity (controlling for winter temperature, and robust to the inclusion of numerous other attributes). An increase in July heat index from one standard deviation below its sample mean to one standard deviation above its sample mean (from 87 °F to 109 °F) is associated with slower growth of 0.5 per cent per year. An increase in relative humidity from one standard deviation below its sample mean to one standard deviation above its sample mean (from 56 per cent to 75 per cent) is associated with slower growth of 0.9 per cent per year. Miami's temperature and humidity in July imply expected annual growth that is 0.7 per cent slower that that of counties with mean heat and humidity.

Finally, population growth is characterized by a negative quadratic partial relationship with the number of rainy days. Increasing the number of rainy days by one standard deviation (25 days) above the mean (94 days) leaves expected population growth essentially unchanged. But increasing rainy days by a second and then a third standard deviation slows growth by 0.3 percentage points and then an additional 0.6 percentage points. For Seattle, with an average of 182 rainy days per year, annual expected population growth is 1.3 percentage points lower than that of a location with mean annual precipitation.

The weather accounts for a very large share of the variation in local population growth rates. The four weather variables just discussed, entered linearly and quadratically, along with annual precipitation entered similarly, can account for 27 percent of the variation in US county population growth from 1970 to 2000. This is only slightly less than is accounted for by dummies for each US state. For metro areas, winter weather alone accounts for 44 percent of the variation in growth from 1950 to 2000.

Results similar to those above hold for a number of nations, for a number of geographies within them, and for a variety of time periods. Similar partial correlations of growth with weather characterize US metro area growth from 1950 to 1980 (Mueser and Graves 1995). In Europe, nice weather has been a major driver of population flows from 1980 to 2000 within countries although not across them (Cheshire and Magrini 2006). And net migration among Japanese prefectures from 1955 to 1990 was negatively correlated with a measure of extreme temperature (Barro and Sala-i-Martin 1995).

The partial correlations strongly suggest that the amenity value of nice weather increased beginning at some point in the intermediate past, via either consumption or production. If the former, such places became inherently more desirable as the marginal utility from nice weather rose relative to the marginal utility of private consumption. If the latter, nice-weather places became more desirable because firms there could pay relatively higher wages.

The quantity framework allows for numerous explanations, many complementary, of the empirical migration to nice weather places. The common element of these explanations is that they posit a change in the valuation of some aspect of weather's amenity contribution, or else a change in the valuation of an amenity correlated with weather. One such explanation is that the approximate sixfold rise in per capita income over the course of the 20th century lowered the marginal utility from the consumption of private goods and services and so increased the quantity of these that households were willing to forgo in order to live in a place with nice weather. Consistent with this consumption amenity explanation, Costa and Kahn (2003), using the compensating differential framework, estimate that a representative household's valuation of enjoying the weather of San Francisco rather than that of Chicago increased more than fivefold between 1970 and 1990.

This rising income explanation for the move to nice weather might intuitively, but incorrectly, be understood to depend on weather's being a luxury good. In fact, it depends only on there being sufficient complementarity between weather and private consumption in the household utility function. Even with a homothetic utility function over private consumption and weather, an increase in income requires a sufficient increase in the valuation of nice weather to dissuade people from moving. More specifically, if the elasticity of substitution between private consumption and weather is exactly 1 (Cobb Douglas), wages and

house service prices can adjust to maintain a spatial equilibrium without any population movement (Rappaport 2009). Essentially a rise in the compensating price of nice weather can exactly cancel an income-driven increase in demand for nice weather. But if instead the elasticity of substitution between weather and private consumption is less than 1, the incomedriven increase in demand is stronger and the larger required offsetting price increase can be supported only if more people move to nice-weather places, thereby driving up housing prices and driving down wages to their general equilibrium values. Conversely, an elasticity of substitution less than 1 will cause the increase in demand for nice weather from increasing incomes to be somewhat weaker. In this case, the required increase in the compensating price is too low to be sustained without some movement away from places with nice weather. Intuitively, a broad, tfp-based increase in wages across all locations can increase the utility cost from not being where wage rates are highest.

A first alternative amenity explanation, based on production, is that the shift to nice weather reflected the movement out of the agriculture and manufacturing sectors. As the share of the labour force employed in agriculture fell from 36 per cent in 1900 to 12 per cent in 1970 to 2 per cent in 2000, the productive amenity contribution of weather to the marginal product of labour averaged over all workers probably decreased greatly. Hence the valuation of weather attributes directly increasing utility *relative* to the valuation of weather attributes conducive to growing would have increased. More recently, as the manufacturing share of employment fell from 25 per cent in 1970 to 14 per cent in 2000, the opportunity cost of moving within the United States from places with perceived less nice weather has probably fallen. On reason is the concentration of heavy manufacturing in the US Midwest, in part due to the proximity of raw materials and notwithstanding winters that are colder and summers that are hotter than many US households desire.

While the declines of agriculture and manufacturing surely contributed to the move to nice weather, they are unlikely to be the main cause. The partial correlation of population growth with nice weather is mostly unaffected by the inclusion of extensive controls for agriculture and other industrial structure. Moreover, the largest part of the move out of agriculture was over by 1970, which is the start date for the growth correlations reported above. Conversely, the move to nice weather began in the 1920s, when manufacturing employment was still growing vigorously.

A second alternative amenity story, based on consumption, is that the move to nice weather reflected the increased mobility and prosperity of the elderly. Rather than the population as a whole, it was primarily the elderly who increased their valuation of nice weather as it became part of their locational choice set. The increase in choice set followed from numerous trends, including the passage of Social Security (pensions for the elderly), increased longevity, and falling transportation and communications costs. Certainly, some warm-weather states such as Florida and Arizona have attracted a disproportionate number of elderly residents from elsewhere. But the strength of the correlation of growth with nice weather is nearly the same for working-age individuals as it is for seniors. Moreover, the move to nice weather began long before the large increases in senior longevity and prosperity.

A third, related, amenity explanation is that for a broad swathe of the US population, mobility costs fell over the course of the 20th century. High moving costs allow for the possibility of rents for those residing in nice-weather places, with the negative compensating differential settling lower (in absolute value) than it would be with free mobility. To the extent that mobility increased – for example, due to falling transportation and communication costs – nice-weather places would have grown disproportionately fast until they reached their free-mobility equilibrium. While this explanation has intuitive appeal, the extent to which mobility increased is unclear. The state-to-state gross migration rate was approximately flat from 1947 to 1975, then fell slightly through 2000.

A fourth alternative consumption amenity explanation for the move to nice weather is that

it was caused by air conditioning. Air conditioning ameliorated the disamenity of hot and humid summer weather, which in turn is correlated with warm winter weather. Hence households no longer needed to be compensated as much to live in hot and humid places, which in turn should have caused a shift in population towards such places. Doubtless there is some truth to this hypothesis, as many of the US metropolitan areas that grew most rapidly from 1950 to 2000 have summer weather that would seem insufferable without air conditioning (for example, the daily high heat index in July for Austin, Texas averages 118 °F). However, the move to nice weather began decades before the widespread diffusion of air conditioning. Moreover, the negative partial correlation of population growth with summer heat and summer humidity is exactly the opposite of what air conditioning is expected to cause. Also tempering the air conditioning explanation is the extremely rapid growth of coastal southern California, where summer weather is relatively mild.

An alternative, nonamenity explanation argues that the correlation of population growth with nice weather is largely a coincidence. Glaeser and Tobio (2008) conclude that the post-war movement to places with nice weather arose from faster productivity growth in nice-weather places accompanied by a high elasticity of housing supply there. The latter was due to some combination of plentiful land and minimal government restrictions on building. The conclusion that weather was not an important driver of the population move to nice weather follows primarily from wage and house price compensating-differential regressions using data from the 1950 through 2000 decennial censuses. These regressions suggest that wages rose quicker but house prices rose slower in places with nice weather than elsewhere. Both of these comparative growth rates suggest that households' relative valuation of nice weather was decreasing over this period.

Certainly, the convergence of productivity in the US South to the national level was an important aspect of the rapid growth of many nice-weather places (Barro and Sala-i-Martin 1991, 1992; Caselli and Coleman 2001). But in the absence of any increase in amenity valuation, the relatively high density and congestion that have come to characterize many nice-weather cities would require productivity there to surpass its level elsewhere, not just converge to it.

Similarly, a relatively elastic housing supply is certainly a necessary condition for the rapid growth that was sustained over 50 years by a number of nice-weather metro areas. In the quantity model described above, the house supply elasticity governs the magnitude of the growth response to a change in amenities. But the impetus for the growth is solely the amenity change. Elastic housing supply, on its own, is not sufficient. Many sparsely populated and declining metro areas throughout the US Midwest and deep South also have plentiful land, light regulation, and in many cases an excess supply of existing buildings.

An additional consideration is the generic unreliability of the compensating differential methodology. The estimated rising wages by Glaeser and Tobio (2008) in nice-weather places may partly reflect an upgrading of unobserved human capital. The increase in the average skills of workers in such metro areas may have been faster than elsewhere. For example, workers who moved to nice-weather places may have had higher skills on average than the skills of workers who already lived there. And slower-than-expected house price growth might reflect that the (negative) compensation for nice weather is being paid, in part, by longer commutes, increased traffic, and other sorts of metro area congestion.

Conclusions

The conclusion that households are shifting towards places with nice weather, at least for the United States, is not very surprising. Indeed, the US business magazine *Forbes* parodied some of the research on the population shift to nice weather with the headline, 'Duh!' (Kellner 2004). Much more important is why households are doing so. The explanations above together suggest that rising incomes caused individuals to

sufficiently increase their valuation of weather as a consumption amenity so as to require a shift in population towards nice weather places. For the increase in valuation to be sufficiently large, weather must have been a complement to private consumption rather than a substitute. The shift towards nice weather was likely reinforced by the change in industrial composition away from agriculture and manufacturing, the increase in productivity throughout the southern United States, the spread of air conditioning, and the increasing mobility and financial security of seniors. Lastly, a high elasticity of housing supply in many nice-weather places implied that the population influxes required to support the increased valuation were quite large.

An important implication of the income result is that valuations of other local consumption amenities are likely to have increased as well. While local governments may be unable to affect their local weather, they may want to consider increasing the supply of other consumption amenities in its place.

A last question is whether the increasing valuation of nice weather and the shift in population towards it are likely to continue. Unambiguously, a continuing increase in income will cause a continuing increase in the valuation of nice weather. For the actual *movement* to nice weather to continue, the increase in valuation must be sufficiently large that it cannot be supported by the existing distribution of population across locations. With sufficient complementarity between weather and private consumption, theory suggests that the move can continue forever, though at a diminishing pace (Rappaport 2009). The increasingly swollen populations of many nice-weather places put downward pressure on their abilities to elastically supply housing and address other sorts of congestion. As housing supply becomes less elastic and other sources of congestion rise, a smaller increase in population can support a given required increase in compensation for local amenities. Consistent with a diminishing shift, decade-by-decade regressions indeed show that the move towards nice weather peaked in the 1970s, and then slowed in each of the 1980s and 1990s.

See Also

- ▶ Climate Change, Economics of
- ▶ Compensating Differentials
- ▶ Housing Supply
- ▶ Systems of Cities
- ▶ Urban Environment and Quality of Life

Acknowledgments The views herein are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of Kansas City or the Federal Reserve System. Thank you to Yi Li for excellent research assistance.

E

Bibliography

Barro, R.J., and X. Sala-I-Martin. 1991. Convergence across states and regions. *Brooking Papers on Economic Activity* 1: 107–182.

Barro, R.J., and X. Sala-I-Martin. 1992. Convergence. *Journal of Political Economy* 100: 223–251.

Barro, R., and X. Sala-I-Martin. 1995. *Economic growth*. New York: McGraw Hill.

Blomquist, G.C., M.C. Berger, and J.P. Hoehn. 1988. New estimates of quality of life in urban areas. *American Economic Review* 78: 89–107.

Caselli, F., and W.J. Coleman II. 2001. The U.S. structural transformation and regional convergence. *Journal of Political Economy* 109: 584–616.

Chen, Y., and S.S. Rosenthal. 2008. Local amenities and life-cycle migration: Do people move for jobs or fun? *Journal of Urban Economics* 64: 519–537.

Cheshire, P., and S. Magrini. 2006. Population growth in European cities: Weather matters – But only nationally. *Regional Studies* 40: 23–37.

Costa, D.L., and M.E. Kahn. 2003. The rising price of nonmarket goods. *American Economic Review* 93: 227–232.

Gabriel, S.A., and S.S. Rosenthal. 2004. Quality of the business environment versus quality of life: Do firms and households like the same cities? *Review of Economics and Statistics* 86: 438–444.

Glaeser, E., and J. Gyourko. 2005. Urban decline and durable housing. *Journal of Political Economy* 113: 345–375.

Glaeser, E.L., and K. Tobio. 2008. The rise of the sunbelt. *Southern Economic Journal* 74: 610–643.

Greenwood, M.J., G.L. Hunt, D.S. Rickman, and G.I. Treyz. 1991. Migration, regional equilibrium, and the estimation of compensating differentials. *American Economic Review* 81: 1382–1390.

Gyourko, J., and J. Tracy. 1991. The structure of local public finance and the quality of life. *Journal of Political Economy* 99: 774–806.

Haurin, D.R. 1980. The regional distribution of population, migration, and climate. *Quarterly Journal of Economics* 95: 293–308.

Hirshleifer, D., and T. Shumway. 2003. Good day sunshine: Stock returns and the weather. *Journal of Finance* 58: 1009–1032.
Kellner, T. 2004. Duh! *Forbes* 173: 50.
Mueser, P.R., and P.E. Graves. 1995. Examining the role of economic opportunity and amenities in explaining population redistribution. *Journal of Urban Economics* 37: 176–200.
Oi, W.Y. 1997. The welfare implications of invention. In *The economics of new goods*, ed. T.F. Bresnahan and R.J. Gordon. Chicago: NBER/University of Chicago Press.
Rappaport, J. 2004. Why are population flows so persistent? *Journal of Urban Economics* 56: 554–580.
Rappaport, J. 2007. Moving to nice weather. *Regional Science and Urban Economics* 37: 375–398.
Rappaport, J. 2008. Consumption amenities and city population density. *Regional Science and Urban Economics* 38: 533–552.
Rappaport, J. 2009. The increasing importance of quality of life. *Journal of Economic Geography* 9: 779–804.
Roback, J. 1982. Wages, rents, and the quality of life. *Journal of Political Economy* 90: 1257–1278.
Rosen, S. 1979. Wage-based indexes of urban quality of life. In *Current issues in urban economics*, ed. P. Miezkowski and M. Straszheim. Baltimore: Johns Hopkins University Press.
Saunders Jr., E.M. 1993. Stock prices and Wall Street weather. *American Economic Review* 83: 1337–1345.
Stover, M.E., and C.L. Leven. 1992. Methodological issues in the determination of the quality of life in urban areas. *Urban Studies* 29: 737–754.

Economic Demography

Allen C. Kelley and Robert M. Schmidt

Abstract

Economic demography is an area of study that examines the determinants and consequences of demographic change, including fertility, mortality, marriage, divorce, location (urbanisation, migration, density), age, gender, ethnicity, population size and population growth. This article reviews and critically evaluates important macroeconomic dimensions of the 'population debates' between the 'optimists' and the 'pessimists' since 1950. It concludes with an examination of demography in the popular 'convergence' growth models of the 1990s.

Keywords

Adult equivalency; Ageing; Agricultural growth and population change; Capital accumulation; Convergence; Demographic drag; Demographic gift; Demographic transition; Diffusion of technology; Diminishing returns; Dismal science; Economic demography; Economic development; Economic growth; Education; Endogenous growth; Fertility; Free rider problem; Human capital; Innovation; Kuznets, S.; Labour productivity; Learning-by-doing; Life expectancy; Life-cycle modelling; Malthus., T. R.; Mortality; Population density; Population growth; Population size; Renewable resources; Research and development; Rule of law; Saving; Simon, J. L.; Subsistence; Technical change

JEL Classifications

J10

Economic demography is an area of study that examines the determinants and consequences of demographic change, including fertility, mortality, marriage, divorce, location (urbanisation, migration, density), age, gender, ethnicity, population size, and population growth. An applied area of research, economic demography draws upon the theoretical and applied fields of economics. For example, the determinants of fertility or migration primarily draw upon microeconomic theory and labour economics, while the consequences of population growth or ageing primarily draw upon macroeconomic theory and development economics.

The field has had a long tradition of controversy, beginning with the publication in 1798 of *An Essay on the Principle of Population* by the Reverend Thomas Malthus. The basic Malthusian model is founded on two propositions: (*a*) population, *when unchecked*, increases at a geometric rate (for example, 1, 2, 4, 8…) and (*b*) food, in contrast, expands at an arithmetic rate (for example, 1, 2, 3, 4…). The result is a population trapped at a meagre standard of living. Short of 'preventive checks' (birth control), population is constrained to live at subsistence by 'positive

checks' (deaths, war, famines and pestilence). In later writings Malthus admitted the possibility of 'moral restraint' that could deter births, primarily through the postponement of marriage. However, he held little hope for a notable attenuation of the 'natural passions' of the working class.

While much of the controversy relating to Malthusianism has focused on the determinants of population growth, a second premise of his model relates to its economic underpinnings: the determinants of agricultural growth. Here Malthus appealed to the historical law of diminishing returns in agriculture. While this proposition engendered relatively little dispute at the time, history has since documented widespread and sometimes notable improvements in agricultural technology. Indeed, food production has represented an engine of growth in many of the areas that Malthus investigated. In some areas today, governments worry about 'excess' food production that depresses prices and farmers' living standards. Unfortunately, the pessimistic food-production predictions, when confronted by rapid population growth, caused economics to be dubbed the 'dismal science'.

The enormous popularity of the Malthusian ideas was the result of several factors: the model's simplicity and its explanation of poverty (the poor failed to exercise moral restraint, ending up with large families); the appeal of the message that subsidising the poor is of questionable efficacy; and the plausibility of the Malthusian argument given the unexpected 'population explosion' revealed by the 1801 census. These and other elements of the 'Malthusian debate' provide a useful taxonomy for organising the present article.

Specifically, we highlight the macroeconomic dimensions of the economic consequences of population growth since 1950. As with the early Malthusian debates, an assessment of the macroeconomic impacts of demographic change on economic production has resulted in an outpouring of research, which has spawned further debate. There are periods when vigorous Malthusian-like alarmism has carried the day; there are periods of counter-challenges; and, since the mid-1980s, there has been a productive 'revisionist' movement. In short, the simplistic Malthusian notion of diminishing returns in production has given way to more informed modelling of economic–demographic interactions. An assessment of the historical evolution of this literature will constitute the bulk of this review and appropriately delimits the scope of our essay since a wide range of important microeconomic themes are taken up in other articles in this dictionary (see ▶ "Fertility in Developing Countries," ▶ "Family Decision Making," ▶ "Marriage and Divorce," and ▶ "Retirement," and multiple articles dealing with the topics of gender, ageing and mortality).

We begin by examining population impacts in one-sector growth models. This leads nicely into a more detailed assessment of factor accumulation, and in particular, the impacts of demography on saving, investment and technological change. This is in turn followed by an analytical description of the evolution of economic–demographic thinking since 1950. Such a perspective exposes many of the key analytical and empirical linkages of interest. The article concludes with an examination of 'convergence modelling', a useful paradigm that exposes the roles of changing demographic structures that take place over the demographic transition.

Theory: Modelling Economic–Demographic Change

One-Sector Growth Models

The aggregate production function constitutes the primary organising device for delineating the impacts of demographic change on economic growth. Within this model, labour productivity depends on the availability of complementary factors of production (land, natural resources, human and physical capital) and technology. If we assume, for convenience, that labour is a constant fraction of population, then population size directly affects aggregate output.

In a production function with constant returns to scale, an increase in population growth will lower the average availability of other factors of production – a 'resource-shallowing' effect, and, through diminishing returns, reduce the growth of worker productivity. Such an adverse

demographic impact can be magnified (or attenuated) if population growth diminishes (raises) the growth rate of complementary factors.

In a standard growth model with factor inputs of labour and capital, and a saving rate and pace of technological change that are exogenous with respect to population growth, demography affects the long-run *level* but not the long-run *growth rate* of output per capita. This is because the capital-shallowing effect of increased population will eventually reduce the capital per worker ratio to a level sufficient to be maintained by a fixed rate of saving. In this case, long-run growth is determined by the pace of technological change. The determinants of the 'fixed' saving rate and pace of technology growth, both considered in more detail below, are central to the analysis.

If one relaxes some of the assumptions of this model, the impact of population growth on per capita output growth can be ambiguous. Negative impacts can arise through diminishing returns, diseconomies of scale, and perhaps savings, while positive impacts can arise through induced technological change, economies of scale, and possibly savings. Most economists believe that adverse capital-shallowing impacts will dominate positive feedback effects, although the magnitude of the demographic impacts may not be all that large.

Saving

Possibly the most investigated linkage of population growth to economic growth has been the impact of demographic change on saving. Two perspectives dominate.

Adult equivalency. Rapid (slow) rates of population growth result in a disproportionate number of children (elderly adults) who consume, but contribute relatively little to, household income. In recognising that these 'dependents' consume less than a working-age adult, the notion of an 'adult equivalent' consumer was born. The financing of an additional child's 'adult-equivalent' consumption has been hypothesised to be out of saving. Such a view, however, has been challenged by consideration of several offsetting alternatives. Specifically, children may (*a*) substitute for other forms of consumption, (*b*) contribute directly to household market and non-market income, (*c*) encourage parents to work more (or less), (*d*) stimulate the amassing (or reduction) of estates, and (*e*) encourage (or discourage) the accumulation of certain types of assets (for example, education or farm implements). The net impact of changing dependency rates on saving is therefore theoretically ambiguous. This is particularly the case if one views human capital as an investment financed in part by households and governments. At any rate, empirical evidence showing negative impacts of youth dependency on saving are found in several studies.

The life-cycle. A second population-saving linkage is based on a life-cycle formulation incorporated into a lifetime household utility function. Specifically, households attempt to even out their lifetime consumption by setting aside earnings during working years to finance consumption by their children as well as for their own retirement. This formulation can yield positive or negative impacts on aggregate saving depending on the relative sizes of the dissaving youth and elderly cohorts. While empirical evidence from life-cycle modelling is mixed, those studies do tend to show linkages between age structure and saving. However, the direction and magnitude of that impact depends upon time and place. (See, for example, Mason 1987; Higgins 1998; and Lee et al. 2001.)

Population-Sensitive Government Spending

Government spending on population-sensitive activities such as schooling (youth) and health (elderly) has been alleged both to reduce saving and to crowd out spending on relatively growth-oriented investments. These two hypotheses constitute the core of Ansley J. Coale and Edgar M. Hoover's (1958) path-breaking study of India. While these premises are appealing, they require qualification. Governments have many options to accommodate population pressures. Indeed, limited empirical evidence (for example, Schultz 1987) has shown that education financing can be met all or in part by (*a*) trade-offs within the public sector, (*b*) reductions in per pupil expenditures, and (*c*) efficiency gains. While the second approach can be expected to reduce the quality of

education (and therefore future productivity), the importance of population pressures on government spending or educational quality is uncertain.

Technological Change: Density, Size and Endogenous Growth

While development economists have for decades harkened the pace of technological change as a (the?) major source of economic growth, most standard growth theory models take the rate of technological change as exogenous. With technological change independent of demographic change, population growth per se will have no impact on the pace of economic growth in long-run equilibrium. By contrast, if technological change is all or in part *embodied* in new investment, then a vintage specification is appropriate whereby new capital is relatively more productive than old. In this set-up, population growth can be economic-growth enhancing by expanding the rate at which technology is incorporated into production. In yet another specification, population growth can directly affect the rate of technological change and/or its form (factor bias). Kenneth J. Arrow (1962) has hypothesised that learning by doing is quickened in an environment of rapid employment growth.

A fourth linkage between technology and demography is found in 'endogenous growth' models that relate the pace of technology directly to population *size.* In particular, the benefits of R&D are assumed to be available to all firms without cost; that is, an R&D industry generates a non-rival stock of knowledge. As a result, if we hold constant the share of resources used for research, an increase in population size advances technological change without limit. This somewhat controversial prediction has been qualified by models that incorporate various firm- or industry-specific constraints on R&D production. Such models typically reduce, but do not eliminate, the positive impacts of population size which, as in the embodiment models above, are manifested largely during the 'transition' to long-run equilibrium.

Evidence on the roles of demographic-technology linkages and growth has been fragmentary and sparse. A pioneering study by Hollis Chenery and Moises Syrquin (1975) draws upon the experience of 101 countries across the income spectrum over the period 1950–70. They find that the structure of development reveals strong and pervasive scale effects (measured by population size) that vary by stage of development. Basically, small countries develop a modern productive industrial structure more slowly and later, while large countries have higher levels of accumulation and (presumably) higher rates of technological change. Although these roles for demography may have been important historically, the impacts plausibly have waned somewhat: (*a*) economies in infrastructure are judged to be substantially exhausted in cities of moderate size; (*b*) specialisation through international trade provides a means of garnering some or many of the benefits of size; and (*c*) scale effects are most prevalent in industries with relatively high capital–labour ratios and such industries are inappropriate to the factor proportions of developing countries.

It is in agriculture where the positive benefits of population size have been most discussed. Higher population densities can lower per unit costs and increase the efficiency of transport, irrigation, extension services, markets and communications (Glover and Simon 1975). Possibly the most cited work is that by Ester Boserup (1965, 1981), who observes that increasingly productive agricultural technologies are made economically attractive in response to higher land densities. While this is probably true, the issue becomes one of identifying the quantitative magnitude of such effects over varying population sizes and in differing institutional settings. One must be cautious in attributing causation. For example, while high population densities may have accounted for a portion of expanded agricultural output in recent decades, in several important Asian countries these densities were sufficiently high decades ago to justify the investments associated with the new technologies. Boserup in more recent writing has been less sanguine about the benefits of population size because densities appropriate to modern technologies in Asia are three to four times the average for Africa and Latin America.

In short, a wide-ranging review of the literature does not provide a strong consensus on the

quantitative linkages between the size and growth of population, on the one hand, and the pace of technological change and economic growth, on the other hand.

The Bottom Line

An evaluation of population growth on economic growth through the filter of formal economic-growth modelling yields limited results: population growth affects the level but not the growth of per capita income in long-run equilibrium. Moreover, the key determinants of long-run growth are saving and technology. Only if these factors depend on demographic change does population matter. This somewhat constraining limitation of growth theory has caused researchers to branch out and explore a host of economic–demographic interactions using less formal paradigms. This blossoming literature has been extensive, lively and sometimes contentious.

Evolution of Population-Impacts Thinking: 1950–90

Four major studies, two by the United Nations (1953, 1973) and two by the National Academy of Sciences (1971, 1986), reveal well the evolution of thinking on population matters over the period 1950–90. Three individual scholars, Coale and Hoover and Simon, also played prominent and important roles. (This section draws on Kelley 2001.)

United Nations, 1953

The 1953 United Nations report, *Determinants and Consequences of Population Trends*, easily represents the most important contribution to population thinking since the writings of Malthus. Unlike Malthus, however, the UN study was balanced and exhaustive both in detail and in coverage. Some 21 linkages between population and the economy were taken up. For example, the impacts of population on the economy can be: (*a*) positive due to economies of scale and organisation; (*b*) negative due to diminishing returns; or (*c*) neutral due to technology and social progress. An evaluation of these and other linkages led to a mildly negative overall assessment that was both cautious and qualified.

The most notable feature of this report was its methodology. More than any major study on population to that time, the UN Report embraced a methodology that would ultimately represent elements of modern-day 'revisionism'. Specifically, the report (*a*) downgraded the importance of population growth's impact on economic growth by placing it on a par with several other determinants of equal or greater impact; (*b*) assessed the consequences of population over a long period of time; and (*c*) emphasised the importance of feedbacks within and between the economic and political systems.

Coale and Hoover, 1958

The next major contribution to the population-impacts literature was provided by Ansley J. Coale and Edgar M. Hoover in their 1958 book *Population Growth and Economic Development in Low-Income Countries*. Based on simulations of a mathematical model calibrated with Indian data, they concluded that India's development would be enhanced by lower population growth. This was due to the hypothesised adverse impacts of population on household saving. It was also proffered that 'unproductive' investments in human capital (such as health and education) would partially displace investments in 'relatively productive' forms (such as machines and factories). Economic growth would diminish in response.

Empirically, the above hypotheses have not been convincingly established. While several studies have exposed negative dependency-rate impacts on saving, there are others that show little or no impact. Overall, the findings are mixed, with a tilt toward supporting the Coale and Hoover formulation. (See section "Saving" above for a discussion of the trade-offs that households can make to maintain saving in response to expanding family size.)

Similarly, there are alternative ways for governments to organise and finance schooling in response to population pressures. Unfortunately, studies of this are limited, although one by T. Paul Schultz (1987) finds no support for the Coale and Hoover (1958) formulation.

National Academy of Sciences, 1971

Arguably the most pessimistic assessment of the consequences of population growth was a study compiled by the National Academy of Sciences (NAS). The panel's final submission, *Rapid Population Growth: Consequences and Policy Implications*, issued in 1971, appeared in two volumes: volume 1, *Summary and Recommendations*, and volume 2, *Research Papers*. Unfortunately, the *Summary* volume appeared to be more political than academic in goal and orientation, and was not faithful to many of the underlying research reports assembled by the panel. Indeed, the *Summary* volume highlighted some 25 alleged negative consequences of population growth, whereas it downplayed or eliminated impacts that could be considered as 'neutral' or 'favourable'. As a result, the *Summary* represents an upper bound on the negative consequences of population growth. (A detailed documentation exposing the somewhat controversial way in which the *Summary* was compiled is provided by Kelley 2001.)

What can be learned from the NAS study? First, given its apparent bias and the lack of a systematic vetting of volume 1 by members of the panel, it is difficult to use that volume, either in full or in part. However, the individual papers are available and they, in total, offer a more balanced treatment. Second, by its own acknowledgment, the study focused on the short run when negative impacts of population change are most likely to prevail. ('We have limited ourselves to relatively short term issues'; 1971, p. vi.) By contrast, 'direct' (short-run) impacts of demographic change are almost always attenuated (and sometimes offset) by 'indirect feedbacks' that occur over longer periods of time. Thus the decision by the NAS panel to focus only on the short-run direct impacts resulted in an overly negative assessment of the consequences of population growth.

Third, economists were underrepresented on both the panel and in providing background reports. This is relevant since economists have substantial faith in the capacity of markets, individuals and institutions to adjust in the face of population pressures. Such adjustments, of course, take time and they are not without cost. Finally, this NAS Report provides a striking example of the difficulty of maintaining objectivity when social science research enters the public policy domain.

United Nations, 1973

In 1973 the United Nations weighed in with an update of its previous seminal work (United Nations 1953). In contrast to the broadly eclectic stance in the earlier report, the new one ended with a mild to moderate negative overall assessment of rapid population growth. The authors were concerned with the ability of agriculture to feed expanding populations (à la Malthus) and the difficulty of offsetting capital shallowing (à la Coale and Hoover). Still, the 1973 Report, whose conclusions are highly qualified, is not alarmist, nor is it all that pessimistic. The reason for this moderate stance was the exceptionally influential empirical finding of Simon Kuznets (1960, pp. 19–20, 63) that notable negative correlations between population growth and per capita output growth were largely absent in the data. Given the strong priors of some contributors to the UN study, a failure to find a negative association in the aggregate data by a scholar with impeccable credentials had a profound impact. Indeed, this singular finding arguably kept the population debate alive for yet another round of assessments in the 1980s.

Revisionism, 1980s and Beyond

The 1980s represented a decade when many of the underlying assumptions and conclusions of earlier studies of population–development interactions were subjected to critical scrutiny. The result was a revisionist rendering that was both surprising and controversial. Specifically, the revisionists downgraded the prominence of population growth as either a major source of, or a constraint on, economic prosperity in the Third World. The basis of this somewhat startling conclusion was the revisionists' methodology that (*a*) assessed the consequences of demographic change over longer periods of time and (*b*) expanded the analysis to take into account *indirect feedbacks* within economic and political systems. In general, empirical assessments of population growth will be smaller

(less negative or less positive) when using the revisionist's methodology than when focusing on the short run and ignoring feedbacks. On net, most revisionists conclude that many, if not most, Third World countries would benefit from slower population growth.

Julian L. Simon, 1981

No one was more important in stimulating the new round of debates in the 1980s than Julian L. Simon, author of *The Ultimate Resource* (1981). This book attracted enormous attention, substantially because of two factors. First, it concluded that population growth would likely provide a *positive* impact on economic development of many developed, and some less developed, countries. Second, the book was accessible, well written, and organised in a 'debating', confrontational style. This included goading and prodding, the setting up and knocking down of straw men, and an examination of albeit popular, but somewhat extreme, anti-natalist positions. Simon's powerful book helped spawn a group of survey articles in the 1980s.

What accounts for Simon's positive assessments? Simon was an early advocate of evaluating the full effects of population over the intermediate to long run. He argued that the negative 'direct' impacts in the short run will probably be moderated, or sometimes overturned, when households, businesses, and/or governments react to changing prices which signal problems of resource scarcity. Two important examples of responses to population pressures can be cited: those relating to technological change and those relating to natural resource scarcity, both highlighted by Simon.

Technological change. Simon hypothesised and attempted to document that the pace of technological change, and its bias, can be stimulated by population pressures. Technological change, in turn, plays a central role in economic growth theory and has been shown in sources-of-growth studies to be a (the?) key to economic growth. Additionally, with respect to population size impacts in general, Simon observes that major social overhead projects (for example, roads, communications and irrigation) have benefited from expanded populations and scale. (For more detail, see section "Technological Change: Density, Size and Endogenous Growth" above.)

Resource depletion. Consider next the impacts of population growth on natural resource depletion. Theoretically an exhaustion of non-renewable resources (for example, coal and minerals) would appear to be inevitable in the long run. However, such a period may be in the indeterminably distant future. By contrast, Simon argued that the most relevant measure of resource scarcity is its price. He prepared many graphs of US non-renewable resource prices (deflated by price indexes in order to focus on 'real' resource trends).

Surprisingly, virtually every resource has experienced a *declining* real price over lengthy periods of time. This means, à la Simon, that resources are becoming *more* abundant over time. It seems that the more resources are used, the more abundant they become! How can this happen? Simple. A rising resource price, due in part to population pressures, triggers several reactions that reduce or even eliminate the apparent resource scarcity. Specifically, in the short run, rising prices encourage an economising of the resource at every level of production and consumption. In the longer run, rising prices stimulate exploration, new methods of extraction and process, and the search for substitutes.

Nevertheless, Simon recognised that market failures, institutional failures, and political factors can all result in less-than-complete adjustments when population and economic development press against resource availabilities. This is particularly the case with renewable resources (such as rain forests, fisheries, the environment, and so forth) where market or institutional failures are pervasive. Without mechanisms to assign and maintain property rights, internalise externalities, and address free rider problems of public and quasi-public goods, government regulation may be required to safeguard renewable resources over time.

National Academy of Sciences, 1986

Some 15 years after the 1971 National Academy Report that highlighted 25 negative consequences of population growth, a new National Academy

Report was released. In contrast to the previous study, the new report was balanced, eclectic and non-alarmist. A careful examination of its bottom line is instructive.

> *On balance*, we reach the *qualitative* conclusion that *slower* population growth would be beneficial to economic development of *most* developing countries. (1986, p. 90; emphasis added)

This qualified assessment reveals key features found in most population assessments in the 1980s. Specifically: (*a*) there are both positive and negative impacts of demographic change (thus 'on balance'); (*b*) the magnitude of the net impacts cannot be determined given current evidence (thus 'qualitative'); (*c*) only the direction of the impact from high to low growth rates can be ascertained (thus 'slower' rather than 'slow'); and (*d*) the net impact varies from country to country. In most cases it will be negative; in some positive; and in others of little impact (thus, 'most developing countries').

What accounts for the dramatic turnaround in the two National Academy assessments? Several factors can be advanced. First, the 1986 report extends the short-run time horizon of the 1971 report to examine individual and institutional responses to the initial impacts of population change: conservation in response to scarcity, substitution of abundant for scarce factors of production, innovation and adoption of technologies to exploit profitable opportunities, and the like. These responses are considered to be pervasive and they are judged to be important. According to the report writers: 'the key [is the] mediating role that human behaviour and human institutions play in the relation between population growth and economic processes' (1986, p. 4).

Second, the 1986 study was assembled almost entirely by economists whose understanding of and faith in markets to induce responses that modify initial direct impacts of population change is far greater than that of other social and biological scientists.

Third, research accumulating over the 15 years between the two reports revealed a need to downgrade: (1) the concern about non-renewable resource exhaustion; (2) the adverse impact of children on the capacity to save, and in turn to undertake productive investments; and (3) the inability to invest in schooling and health facilities.

Finally, the 1986 Report upgrades the concern about population impacts on *renewable* natural resources (such as fishing areas and rain forests) where property rights are difficult to assign and maintain. Overuse can result. It is recognised that the problems of overuse are not solely due to population growth per se, but rather institutional failure. Cutting population growth by one half, or even to zero, would not solve the problem. Rather it would slow the process and postpone the date of resource exhaustion. Government policies are needed to account for negative externalities and market failure. Slowing population growth provides time for institutional response.

New Paradigms for Modelling Demography's Role in Economic Growth: 1990 and Beyond

As noted previously, Kuznets's empirical finding of an absence of notable negative correlations between population growth and per capita output growth influenced the population debate throughout the 1970s and 1980s. Simple correlations stimulated research during the 1990s as well. This time, however, statistically significant negative correlations during the 1980s drove the discussion. Interestingly, economic–demographic modelling continued in the 'revisionist' vein, incorporating positive and negative as well as short- and long-run influences into an economic growth model. The modelling challenge remains one of accommodating correlations that can be negative, positive or insignificant depending upon time and place.

Convergence Growth Models: A Framework for Assessing Demography's Impact

Renewed interest in modelling the impacts of demographic change on economic growth coincided with the emergence in the economic growth literature of the 'technology gap' or 'convergence' model. This model, formulated initially

by Barro and Sala-i-Martin (1991), has been used widely to explore many hypothesised influences on economic growth, including openness to trade, form of government, and the rule of law. Since this type of modelling highlights the dynamics of the adjustment process, it is particularly relevant to examining the impacts of major shifts in the population's age distribution associated with birth and death rates that change systematically over the demographic transition. As a result, economic demographers have employed convergence paradigms to explore demographic–economic interactions.

Briefly stated, convergence models focus on the pace at which countries move from their current level of labour productivity to their long-run or steady-state level of labour productivity. The model assumes that all countries converge at the same rate from their current to their long-run levels (which can vary across countries and over time). The greater the productivity gap, the greater are the gaps of physical capital, human capital and technical efficiency from their long-run levels. Large gaps allow for 'catching up' through (physical and human) capital accumulation, and technology creation and diffusion across countries and over time. Indeed, many empirical studies indicate that growth rates do slow down as a country approaches its long-run productivity level, especially those studies that provide for country- and period-specific conditions that influence the long-run level of labour productivity.

Since long-run labour productivity is unobservable, empirical implementations of the model substitute a vector of 'conditioning' variables thought to influence long-run labour productivity. The actual specification of these conditioning variables varies notably. Consider two of their many representations. The first, by Barro (1997), highlights inflation, government consumption ratios, the rule of law, the form of the political system, terms of trade, human capital, the total fertility rate, and life expectancy at birth (a proxy for health). The second formulation, by Bloom and Williamson (1998), highlights two categories of growth-rate determinants: economic structure variables (natural resources, schooling, access to ports, location in the tropics, whether landlocked, and extent of coastline); and economic and political policies (openness to trade, quality of institutions, and government savings share of GDP). Clearly there are many defensible perspectives on variable choice, and much is yet to be learned about the appropriate configuration of conditioning variables that influence long-run productivity levels.

Alternative Demographic Renderings Within a Convergence Framework

The 1990s witnessed attempts by various researchers to model demography in a manner that accommodates both the insignificant correlations of the 1960s and 1970s as well as the significant negative correlations of the 1980s and 1990s. Three different approaches are described here. All three employ a convergence-type growth model and all employ a broad set of countries spanning the income spectrum.

Modelling through aggregate measures of fertility and mortality. Barro (1997) includes two demographic aggregate measures among his list of conditioning variables, the total fertility rate (TFR) and life expectancy. Barro's formulation thus has demography impacting the long-run equilibrium level of per capita income. The TFR captures, for example, the adverse capital-shallowing impact of more rapid population growth as well as the resource opportunity costs of bringing up children. Furthermore, while Barro treats life expectancy as a human capital proxy for health, demographers consider it to be a demographic variable. Both are statistically significant, with a higher TFR inhibiting, and longer life expectancy enhancing economic growth.

Modelling through population growth components. Kelley and Schmidt (1995) decompose population growth by examining two components (births and deaths) and by modelling their contemporaneous and lagged impacts. This approach allows for disparate impacts of fertility and mortality as well as negative short-run effects (costs of high birth and death rates) and positive long-run effects (favourable impacts of past births on current labour force growth and declining mortality). Consistent with Kuznets's earlier work, they

found an absence of a net demographic impact on economic growth in the 1960s and 1970 – the separate impacts of births and deaths are notable but offsetting. Consistent with empirical work of the early 1990s, they found negative impacts throughout the 1980s. These negative correlations were in part the result of (*a*) rising short-run costs of high birth rates, (*b*) declining benefits of mortality reduction, and (*c*) insufficient labour force entry from past births to offset these increased costs.

Modelling through differential age-structure growth. In a series of papers beginning in the late 1990s, several Harvard economists argued for a demographic rendering that incorporates not only population growth but also labour growth (see, for example, Bloom and Williamson 1998; and Bloom et al. 2000). They note that, while theorists conceptualise the economic growth process in labour productivity terms, empirical growth models are generally specified in per capita terms. This makes no difference when population and labour grow at the same rate, but does when they grow at different rates.

The authors argue that the post-war period was exactly such a time since during that period demographic transitions took place in different countries at different times and at different paces. At various stages of the demographic transition, the population and working ages (used within this framework as a proxy for labour) can grow at very different rates. In a predictable pattern, the population initially grows faster, then slower, and then faster than the working-aged population during the transition from a high-fertility, high-mortality to a low-fertility, low-mortality demographic steady-state equilibrium. (For an historical evolution of economic, sociological, and biological factors during the demographic transition, see R.A. Easterlin 1978.)

Without allowing for differential growth rates of the population and working ages, demographic coefficient estimates (mainly population growth) will be biased. In that case the population–growth coefficient captures net demographic impacts that can be positive, negative, or neutral, depending upon time and place. Bloom and Williamson (1998) demonstrate this point for a broad cross-section of countries over the period 1965–90 in a convergence model that also includes life expectancy as a human capital variable. Consistent with some studies, their simple demographic rendering results in a positive but insignificant coefficient for the population growth rate. When supplemented by the working-age growth rate, however, that coefficient turns negative and the coefficient for the working-age growth rate is positive, both statistically significant.

Effectively, the Harvard economists append an accounting structure to translate labour productivity impacts into per capita terms. The resulting demographic specification is elegant in its simplicity, incorporating only two demographic variables that have unambiguous predicted coefficient values of – 1 (for population rate of growth, Ngr) and +1 (for working-age population rate of growth, WAgr) when used to expose demography's impact on income growth per capita relative to income growth per working-age population. In that context, demography exerts its primary impact on the pace at which the long-run equilibrium is reached (Bloom and Williamson 1998, p. 419) rather than on the long-run equilibrium level of productivity.

This is an intriguing specification. The interpretation is clear: if labour force growth exceeds population growth, then the rate of per capita income growth is boosted by demography. The Harvard economists label this phenomenon the 'demographic gift' that may be reaped for several decades after the onset of fertility decline as new labour force entrants from earlier large birth cohorts outpace fertility. The 'gift' was large throughout the 1965–90 period for Japan and other Asian Tigers because of the early and rapid pace of their demographic transition. Of course, the converse of the 'gift' began to be felt in the 1990s as new labour force entry from smaller birth cohorts was outpaced by labour force exit of the ageing population. The model predicts productivity outpacing per capita income growth over several decades into the future in these Asian (and other) countries.

Note that the qualitative predictions are based on theoretically determined coefficients on WAgr and Ngr of +1 and −1, respectively. To the extent

that estimated coefficients deviate from +1 and −1, WAgr and Ngr play an additional role in the determination of the long-run productivity level. The Harvard studies provide some guidance in this area. In their earlier study, Bloom and Williamson (1998) estimate coefficients that differ significantly from +1 and −1. However, in a later study that further elucidates the accounting, Bloom et al. (2000) find no significant difference from those values. If that is the case, then the model at once makes an important contribution and is somewhat narrower than many in the literature which admit both short-run and long-run impacts of demographic change as a part of the theoretical structure. Yet modelling demography in growth equations tends to be both imprecise and ad hoc. In contrast, the Bloom and Williamson model is relatively clear in interpretation, and it targets the shorter-run impacts that are of primary interest to policymakers.

The Bottom Line

Bloom and Williamson (1998) estimated that as much as one-third of the average per capita income growth rate in East Asian countries over the period 1965–90 is explained by population dynamics. Kelley and Schmidt (2001) evaluated eight distinct demographic renderings within a convergence model using a consistent set of conditioning variables – those described above for Barro's variant. Among others, these renderings included Barro's TFR; a 'naive' variant predating the 1990s work that simply includes Ngr; a 'components' model (contemporaneous and lagged birth rates and the death rate: Kelley and Schmidt 2001); two variants of the Harvard transitions framework; and demographic extensions to several variants.

Kelley and Schmidt (2001) find that on average, across all eight demographic formulations and over their full 86-country sample (covering the full income spectrum), approximately 21 per cent of the combined impacts on change in the per capita income growth rate is accounted for by changes in the demographic variables in the various models. What is striking about this result is that the 21 per cent is fairly stable across all eight demographic renderings, from one that is quite simplistic (Ngr only) to those that incorporate short-, intermediate- and long-term population effects. On the one hand, this should not be terribly surprising because of the interconnectedness of all of the demographic measures. On the other hand, while population matters, it is still important to determine why.

Although there is an emerging consensus that the magnitude of the impacts of population growth have been sizeable (for example, 21 per cent globally and as much as 33 per cent in East Asia), the reasons why this is the case are still both contestable and not well understood. Are the demographic determinants primarily longer-run impacts, or are they mainly shorter-run transitional dynamics that are diminishing? Will the so-called 'demographic gift' of these dynamics in the past reveal themselves as a 'demographic drag' in the future, deriving from reduced fertility, slow population growth and ageing? Or will a new mechanism reveal itself? For example, (*a*) will future modelling better expose the components of labour force change (for example, utilisation rates, age- and/or gender-specific participation rates); and (*b*) will fertility and mortality be endogenously specified to better reveal the dynamics of the demographic transition about which the field of economic demography has much to say? Whatever the outcome, the stage is set for another round of research, pinning down the results of the past with the goal of understanding the future.

See Also

- Family Decision Making
- Fertility in Developing Countries
- Marriage and Divorce
- Retirement

Bibliography

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29: 155–173.

Barro, R.J. 1997. *Determinants of economic growth: A cross-country empirical study*. Cambridge, MA: MIT Press.

Barro, R.J., and X. Sala-i-Martin. 1991. Convergence. *Journal of Political Economy* 100: 223–251.
Birdsall, N., A.C. Kelley, and S. Sinding (eds.). 2001. *Demography matters: Population change, economic growth and poverty in the developing world*. Oxford: Oxford University Press.
Bloom, D.E., and J.G. Williamson. 1998. Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* 12: 419–455.
Bloom, D.E., D. Canning, and P. Malaney. 2000. Demographic change and economic growth in Asia. *Population and Development Review* 26: 257–1990.
Boserup, E. 1965. *Conditions of agricultural growth*. Chicago: Aldine.
Boserup, E. 1981. *Population and technological change*. Chicago: University of Chicago Press.
Chenery, H., and M. Syrquin. 1975. *Patterns of development: 1950–1970*. Oxford: Oxford University Press.
Coale, A.J., and E.M. Hoover. 1958. *Population growth and economic development in low-income countries*. Princeton: Princeton University Press.
Easterlin, R.A. 1978. The economics and sociology of fertility: A synthesis. In *Historical studies of changing fertility*, ed. C. Tilly. Princeton: Princeton University Press.
Glover, D.R., and J.L. Simon. 1975. The effect of population density on infrastructure: The case of road building. *Economic Development and Cultural Change* 23: 453–468.
Higgins, M. 1998. Demography, national savings, and international capital flows. *International Economic Review* 39: 343–369.
Kelley, A.C. 1988. Economic consequences of population change in the Third World. *Journal of Economic Literature* 26: 1685–1728.
Kelley, A.C. 2001. The population debate in historical perspective: Revisionism revisited. In *Population matters: Demographic change, economic growth, and poverty in the developing world*, ed. N. Birdsall, A. Kelley, and S. Sinding. Oxford: Oxford University Press.
Kelley, A.C., and R.M. Schmidt. 1995. Aggregate population and economic growth correlations: The role of the components of demographic change. *Demography* 32: 543–555.
Kelley, A.C., and R.M. Schmidt. 2001. Economic and demographic change: A synthesis of models, findings and perspectives. In *Population Matters : Demographic Change, Economic Growth, and Poverty in the Developing World*, ed. N. Birdsall, A.C. Kelley, and S. Sinding. Oxford: Oxford University Press.
Kuznets, S. 1960. Population change and aggregate output. In *Demographic and economic change in developed countries*, National Bureau of Economic Research. Princeton: Princeton University Press.
Lee, R.D., A. Mason, and T. Miller. 2001. Saving, wealth and population. In *Population matters: Demographic change, economic growth, and poverty in the developing world*, ed. N. Birdsall, A.C. Kelley, and S. Sinding. Oxford: Oxford University Press.
Malthus, T.R. 1798. *An essay on the principle of population*, , 1970. Harmondsworth: Penguin.
Mason, A. 1987. National saving rates and population growth: A new model and new evidence. In *Population growth and economic development: Issues and evidence*, ed. D.G. Johnson and R.D. Lee. Madison: University of Wisconsin Press.
National Academy of Sciences. 1971. *Rapid population growth: Consequences and policy implications, vol. 1: Summary and recommendations; vol. 2: Research papers*. Baltimore: Johns Hopkins University Press.
National Academy of Sciences. 1986. *Population growth and economic development: Policy questions*. Washington, DC: National Research Council.
Schultz, T.P. 1987. Schooling expenditures and enrollments 1960–1980: The effects on income, prices and population growth. In *Population growth and economic development issues and evidence*, ed. D. Gale Johnson and R.D. Lee. Madison: University of Wisconsin Press.
Simon, J.L. 1981. *The ultimate resource*. Princeton: Princeton University Press.
Simon, J.L. 1996. *The ultimate resource 2*. Princeton: Princeton University Press.
Srinivasan, T.N. 1988. Modeling growth and economic development. *Journal of Policy Modeling* 10: 7–28.
United Nations. 1953. *The determinants and consequences of population trends*. New York: United Nations.
United Nations. 1973. *The determinants and consequences of population trends*. New York: United Nations.

Economic Development and the Environment

Ian Coxhead

Abstract

Economic development in low-income economies is initially highly resource-intensive. Resource depletion and pollution damage is often estimated to reduce 'real' GDP growth by between one and two per cent per year. Growth and structural change alter the environment–development nexus in nonlinear fashion. Policy reforms, global market integration, and institutional development all alter the propensity for growth to generate environmental damage. The emergence of new trade patterns among developing countries has created

new challenges in the measurement and analysis of development–environment interactions. Larger developing economies are now emerging as major sources of emissions that contribute to global climate change.

Keywords

Biodiversity; Comparative advantage; Conservation; Economic development and the environment; Environmental economics; Environmental Kuznets curve; Greenhouse gas emissions; Growth and international trade; Heckscher–Ohlin trade theory; Import substitution; Income effects; Natural resources; Non-use amenities; North–South economic relations; Pollution; Pollution haven hypothesis; Poverty alleviation; Ricardian trade theory; Social norms; Structural change; Sustainable development

JEL Classifications

O10

Economic development depends on sustained per capita income growth and entails dramatic changes in production structure. In low-income economies, growth typically stimulates markets and promotes the evolution of institutions that constrain behaviour according to social norms. The expansion of trade in relation to GDP is another common accompaniment to growth. Each of these has effects on 'the environment', which in a developing-country setting refers not only to phenomena such as water and air quality but also, importantly, to natural resource stocks such as forests, fisheries and soils.

Conversely, changes in environmental quality, including resource stock drawdowns, may affect economic development in a dynamic interaction. This feedback is hard to quantify; however, the World Bank's *World Development Indicators* series now includes 'adjusted' national accounts data reporting GDP and savings net of the implied value of resource depletion and environmental damage (Bolt et al. 2002). These indicate that environmental damage can reduce GDP growth by as much as one to two per cent per year. On a broader scale, growth of large low-income economies like China and India is beginning to have ramifications not only for their own environmental conditions, but also for the global environment through transboundary pollution spillovers and greenhouse gas (GHG) emissions.

The welfare of the poor in low-income countries is intimately linked to their access to environmental assets, and especially to the natural resource base. Despite this, the central concerns of environmental and resource economics – the economic costs of pollution and natural resource depletion – have only recently begun to be linked to models of economic development. Publication of the so-called Brundtland Report (WCED 1987) was a watershed event; since then, 'no account of economic development would be regarded as adequate if the environmental-resource base were absent from it' (Dasgupta and Mäler 1995, p. 2734).

Growth in low-income economies is inevitably associated with higher resource demands and increased pollution intensity per unit of income generated. Other things equal, more economic activity generates more environmental damage monotonically through a scale effect. The relationship may be nonlinear, however. As income grows, environmental damage per unit of additional income may initially rise, then decline. This conjecture, known as the environmental Kuznets curve (EKC), posits that scale effects dominate all other influences on the growth–environment relationship at low income levels, but that, as incomes rise, changes in the composition of production, technological improvements, and income-elastic preferences for conservation and a cleaner environment become more influential (Grossman and Krueger 1993). Institutional and legal constraints on pollution and resource depletion, initially so weak as to create a form of open access for polluters and resource depleters, may also evolve or be applied with greater vigour as incomes increase, whether due to income effects or to increased recognition of limits to growth imposed by pollution and resource scarcity (Stokey 1998). Despite the heuristic value of EKC, however, empirical tests in low-income economies are plagued by data and measurement

problems. Most notably, there is no robust evidence of an EKC for resource-depleting activities such as deforestation.

Changes in production structure and factor demands are also inherent to development. The most prominent manifestation of structural change in low-income countries is the relative decline of agricultural and resource sectors as contributors to GDP and employment. This has clear environmental implications when the majority of the population is initially dependent on the natural resource base. In capital-scarce economies, forest and land conversion for agriculture and the exploitation of fisheries and other resource stocks are standard strategies for increasing labour productivity and generating surpluses. Accordingly, early stages of development are characterized by rapid resource depletion – most visibly in the form of tropical deforestation. Such processes are abetted by conditions of open access (Barbier 2005).

Whether the depletion rate eventually slows – a prerequisite for sustainable development – depends largely on the extent to which surpluses are used to build capacity in secondary and tertiary industries making more intensive use of reproducible resources such as labour, technology and human capital. In this way, the central story of structural change in low-income economies is intimately linked to the evolution of demands on the environmental and natural resource base. Sustained growth leads to a relative reduction in dependence on natural resources, and thus makes it easier for society to agree to promote conservation, biodiversity retention and non-use amenities. Conversely, macroeconomic failures, often in combination with rapid population growth, high transactions costs and market failures, can lead low-income economies into unsustainable cycles of poverty, resource over-exploitation, and institutional failure.

Trade is another influential source of structural change. Early development policies stressing import substitution and de-emphasizing trade have, in most countries, been supplanted by greater outward orientation. Trade-to-GDP ratios have risen and domestic prices have tended to converge on world market prices, thus altering domestic production and investment incentives. With the exception of resource-poor East Asian countries like Korea and Taiwan, the pursuit of comparative advantage in low-income countries initially means expanded exports of tropical agriculture, forestry and fisheries and of resource-based semi-manufactures such as sawnwood. Both the growth of global demand and the pro-trade effects of policy reforms encourage accelerated resource drawdowns; unless property rights and externalities are adequately dealt with, these are likely to occur at socially excessive rates (Coxhead and Jayasuriya 2003). A related idea known as the pollution haven hypothesis posits that weak environmental laws and unresolved externalities may lead developing countries to specialize in pollution-intensive industrial activities (Copeland and Taylor 1994).

Whereas early policy advice to developing countries typically stressed the desirability of exploiting resource wealth to create jobs and earn foreign exchange, contemporary concerns about exhaustibility and the integrity of ecological systems have led to more cautious counsel and an emphasis on sustainable development. Such advice, however, is often difficult to implement as policy in the face of pressures to promote growth and alleviate poverty in the current generation.

New issues in the development–environment relationship continue to emerge as economies grow and become more globalized. Traditionally, trade-environment analyses used Ricardian or Heckscher–Ohlin models of North–South interactions in which welfare growth in resource-abundant South is contingent on trade with industrialized North and on domestic externalities or market failures (for example, Chichilnisky 1994). However, South–South trade – or, in the case of China's emergence as a major market for resource exports from Asia, Africa, and Latin America, 'East–South' trade – is now growing much faster than trade of the North–South type. South–South trade is a form of internationally fragmented production in which primary products or semi-manufactures are exported from one low-income country to another to be used in production of final goods. The latter low-income economy thus moves to 'clean' growth based on labour-intensive

manufactures, while growth in the former becomes more resource-intensive. Countries in the South may have comparative advantage in either clean or dirty goods – or both. Conventional models and measures for evaluating environmental costs of growth must be adapted to such new modalities.

Other new trends reflect the growing global influence of large developing economies. In poor countries, about 50 per cent of carbon dioxide emissions (the primary sources of GHGs) comes from land conversion. But total emissions increase rapidly with energy demands driven by growth, urbanization and industrialization. According to the International Energy Agency, China accounted for 13 per cent of global energy-related CO_2 emissions in 2006, and is expected to overtake the USA as the largest CO_2 source by 2009; India is now following a similar path (IEA 2006). Under the 1997 Kyoto Protocol, these economies are not required to limit GHG emissions. But, even if they do take major steps to limit pollution intensity, scale effects of their growth will ensure that global pollution externalities will continue to expand for the foreseeable future. In turn, concerns over the global environmental consequences of growth in low-income countries will find increasingly forceful expression in international negotiations not only on the environment but also on trade and other forms of international integration.

See Also

- ▶ Climate Change, Economics of
- ▶ Environmental Economics
- ▶ Environmental Kuznets Curve
- ▶ Poverty Alleviation Programmes
- ▶ Sustainability

Bibliography

Barbier, E.B. 2005. *Natural resources and economic development*. Cambridge: Cambridge University Press.

Bolt, K., M. Matete, and M. Clemens. 2002. *Manual for calculating adjusted net savings*. Environment Department, World Bank: Mimeo.

Chichilnisky, G. 1994. North–South trade and the global environment. *American Economic Review* 84: 851–874.

Copeland, B.R., and M.S. Taylor. 1994. North–South trade and the environment. *Quarterly Journal of Economics* 109: 755–787.

Coxhead, I., and S.K. Jayasuriya. 2003. *The open economy and the environment: Development, trade and resources in Asia*. Cheltenham/Northampton: Edward Elgar.

Dasgupta, P., and K.-G. Mäler. 1995. Poverty, institutions and the natural resource base. In *Handbook of development economics*, ed. J. Behrman and T.N. Srinivasan, Vol. 3A. Amsterdam: North-Holland.

Grossman, G.M., and A.B.. Krueger. 1993. The environmental impacts of a North American Free Trade Agreement. In *The US-Mexico free trade agreement*, ed. P. Garber. Cambridge: MIT Press.

IEA (International Energy Agency). 2006. *World energy outlook 2006*. Paris: IEA.

Stokey, N. 1998. Are there limits to growth? *International Economic Review* 39: 1–31.

WCED (World Commission on Environment and Development). 1987. *Our common future*. Oxford/New York: Oxford University Press.

Economic Epidemiology

Tomas J. Philipson

Abstract

The economic analysis of epidemiological issues has different implications for disease and its optimal control from those of traditional analysis of such issues. It views undesirable disease occurrence as the result of self-interested behaviour in the presence of constraints. Unlike with methods used in public health, the effects and desirability of disease-reducing public interventions are then evaluated in terms of how they improve the private behaviour essential to controlling disease. Economic epidemiology has been applied to a wide range of topics, including infectious diseases such as AIDS, and also to non-infectious behaviour such as smoking, obesity, and crime.

Keywords

Economic epidemiology; Health subsidy; Morbidity; Mortality; Programme evaluation; Public health; Rational epidemics

JEL Classifications
I1

The fast-growing literature on the economic analysis of epidemiological issues (see Philipson 2000, for a review) delivers very different implications about disease occurrence and its optimal control from those of traditional analysis of the same issues in the field of public health. At the risk of vastly oversimplifying the positive component of the public health approach, the traditional analysis comprises empirical methods and analysis aimed at identifying and quantifying the effects of 'risk factors' on health outcomes. These factors are typically defined as covariates that negatively affect the measured health outcomes – for example, the effects of smoking on lung cancer or the effects of obesity on heart disease. Thereafter, the normative component of the public health approach is concerned with attempts to reduce the measured risk factors, whether through private or public intervention, and to thereby improve health outcomes.

This approach drastically differs from that of economic epidemiology, which attempts to explain undesirable disease occurrence as the result of self-interested behaviour in the presence of constraints. The effects and desirability of disease- reducing public interventions are then evaluated in terms of how they improve the private behaviour essential to controlling disease in the first place. In some sense, the public health approach aims to improve health, whereas the economic approach aims to improve economic efficiency, even if that does not necessarily improve health. Just as closing highways would improve health but impair economic efficiency, the two approaches often clash in desired interventions. The public health approach, therefore, more often favours public intervention, and sometimes simply assumes that the existence of a health problem is sufficient cause for intervention, potentially because it lacks a theory about how private incentives affect the observed level of disease across time and populations.

Economic Epidemiology and Infectious Disease

Infectious diseases cause roughly one-third of all deaths worldwide and represent the primary cause of mortality in the world. Historically, the share of worldwide mortality due to infectious diseases has been even greater, although data tend to be less reliable for earlier periods. Morbidity and mortality from infectious diseases such as tuberculosis, malaria and acute respiratory infection have always been at the forefront of public policy in developing countries, where infectious diseases accounted for nearly one-half of mortality in the 1990s.

Worldwide concern about infectious disease has received renewed interest in public policy discussions given the disastrous impacts of HIV/AIDS and the potential threat of bird flu. Like most communicable diseases, especially those that are potentially fatal, HIV has incited an extensive governmental response, consisting of regulatory measures, subsidies for research, education, treatment, testing and counselling. Here we review the main contributions of economic epidemiology in predicting both the short- and the long-run behaviour of infectious disease, as well as the effects and desirability of public health interventions that attempt to reduce such disease.

Philipson and Posner (1993) provide the first systematic analysis of rational infectious disease epidemics in the context of AIDS. Kremer (1996) analyses the effects of a reduction in the number of one's sexual partners on the growth of disease. The predictions of such models rely crucially on the prevalence elasticity of private demand for prevention against disease, that is, the degree to which prevention increases in response to disease occurrence. Prevalence-elastic behaviour has different implications for the susceptibility to infection than standard epidemiological models of disease occurrence as discussed in Philipson (1995). Evidence of the degree of prevalence-elastic demand is discussed in Ahituv et al. (1996) and Auld (2003, 2006). Oster (2006) attempts to explain the lack of

prevalence-elastic demand in Africa by the competing risks that lower the demand for prevention in that part of the world. Lakdawalla et al. (2006) provide evidence that demand is sensitive to overall risk, both in terms of prevalence and the cost of infection as when reduced by new medical technologies.

This type of prevalence-elastic behaviour has two major implications. First, growth of infectious disease is self-limiting because it induces preventive behaviour. Second, since the decline of a disease discourages prevention, initially successful public health efforts actually make it progressively harder to eradicate infectious diseases. Geoffard and Philipson (1996) discuss a very general result concerning the inability of private markets to eradicate disease when demand is prevalence-elastic because a disappearing disease implies less prevention. Barrett (2003) and Barrett and Hoel (2004) also analyses the implications of economic efficiency for optimal eradication. See also Gersovitz and Hammer (2003, 2004, 2005).

Regarding the value of public health interventions, Mechoulan (2004) analyses the prevalence and efficiency implications of HIV testing. Geoffard and Philipson (1996) argue that eradication is never Pareto optimal when only the current generation is considered. However, the missing market is dynamic: future generations cannot pay vaccine producers for the benefit they derive from the producers' product. Brito et al. (1991) analyse the nonstandard efficiency implications of mandatory vaccinations.

Moreover, the prevalence elasticity of demand lowers the price elasticity of demand, which implies that Pigouvian-style subsidies to stimulate prevention may have only limited success. This occurs because demand rises among those who are subsidized and falls among those who are not – in the extreme case, total demand is inelastic to subsidies. In addition, prevalence competes with public interventions in inducing protective activity, which makes the timing of the public intervention a crucial factor in determining its economic efficiency. If the subsidy is not prompt enough, the growth in prevalence will have already induced protection.

A growing literature examines the optimal control of infectious diseases in the presence of antibiotic resistance (see, for example, Laxminarayan and Brown 2001; Laxminarayan and Weitzman 2002; Laxminarayan 2002; and Horowitz and Moehring 2004). The standard, positive external effect of treating more individuals with an infectious disease is partly or fully offset by the negative external effect induced by increased antibiotic resistance. The R&D problem induced by external consumption effects such as antibiotic resistance is discussed in Philipson et al. (2006).

Economic epidemiology has also considered the welfare losses induced by disease, the welfare effects of R&D in developing new methods of prevention and treatment (Philipson 1995), and how these contrast with cost-of-illness studies of disease burden.

Spread of Economic Epidemiology to Other Fields

Several other topics have grown out of this more systematic analysis of infectious disease by economists. One strand is the analysis of public health-related issues such as obesity (Philipson and Posner 2003; Lakdawalla et al. 2005). The addictive aspect of obesity is analysed by Cawley (1999). Empirical studies explaining the observed growth in obesity, whether it includes a rise in caloric intake or fall in caloric expenditure, include Cutler et al. (2003). Chou et al. (2004) and Rashad and Grossman (2004) analyse the co-variation between the growth of obesity and smoking and fast-food establishments in the United States. The important and rich set of issues raised by growth in obesity promises a useful role for economic analysis.

Another area in which economic analysis of epidemiological issues has emerged is the economic analysis of clinical trials (see, for example, Philipson and DeSimone 1997; Philipson and Hedges 1998; Malani 2006). This literature deals with the non- traditional aspects of programme evaluation that are unique to clinical trials – for example, the blinding of subjects. Economic

analysis of clinical trials differs from bio-statistical analysis in that subjects are assumed to act in their best interest rather than be passively observed.

The stark difference between economic explanations of disease occurrence on the one hand and the evaluation of public interventions aimed at limiting disease on the other implies that economics may have a very useful role to play in understanding these issues.

See Also

► Health Economics

Bibliography

Ahituv, A., J. Hotz, and T. Philipson. 1996. Is AIDS self-limiting? Evidence on the prevalence elasticity of the demand for condoms. *Journal of Human Resources* 31: 869–898.

Auld, M.C. 2003. Choices, beliefs, and infectious disease dynamics. *Journal of Health Economics* 22: 361–377.

Auld, M.C. 2006. Estimating behavioral response to the AIDS epidemic. *Contributions to Economic Analysis and Policy,* 5(1), Article 1.

Barrett, S. 2003. Global disease eradication. *Journal of the European Economic Association* 1: 591–600.

Barrett, S., and M. Hoel. 2004. Optimal disease eradication. Working Paper No. 604, FEEM.

Brito, D., E. Sheshinski, and M. Intriligator. 1991. Externalities and compulsory vaccinations. *Journal of Public Economics* 45: 69–90.

Cawley, J. 1999. Obesity and addiction. Ph.D. dissertation, Department of Economics, University of Chicago.

Chou, S.Y., M. Grossman, and H. Saffer. 2004. An economic analysis of adult obesity, results from the behavioral risk factor surveillance system. *Journal of health Economics* 23: 565–587.

Cutler, D.M., E.L. Glaeser, and J.M. Shapiro. 2003. Why have Americans become more obese? *Journal of Economic Perspectives* 17(3): 93–118.

Geoffard, P.Y., and T. Philipson. 1996. Rational epidemics and their public control. *International Economic Review* 37: 603–624.

Gersovitz, M., and J.S. Hammer. 2003. Infectious diseases, public policy, and the marriage of economics and epidemiology. *The World Bank Research Observer* 18: 129–157.

Gersovitz, M., and J.S. Hammer. 2004. The economical control of infectious diseases. *Economic Journal* 114: 1–27.

Gersovitz, M., and J.S. Hammer. 2005. Tax/subsidy policies toward vector-borne infectious diseases. *Journal of Public Economics* 89: 647–674.

Grossman, M., and I. Rashad. 2004. The economics of obesity. *Public Interest* 156: 104–112.

Horowitz, B.J., and B.H. Moehring. 2004. How property rights and patents affect antibiotic resistance. *Health Economics* 13: 575–583.

Kremer, M. 1996. Integrating behavioral choice into epidemiological models of the AIDS epidemic. *Quarterly Journal of Economics* 111: 549–573.

Lakdawalla, D., T. Philipson, and J. Bhattacharya. 2005. Welfare enhancing technological change and the growth of obesity. *American Economic Review* 95: 253–258.

Lakdawalla, D., N. Sood, and D. Goldman. 2006. HIV breakthroughs and risky sexual behavior. *Quarterly Journal of Economics* 121: 1063–1102.

Laxminarayan, R. 2002. How broad should the scope of antibiotics patents be? *American Journal of Agricultural Economics* 84: 1287–1292.

Laxminarayan, R., and G.M. Brown. 2001. Economics of antibiotics resistance: A theory of optimal use. *Journal of Environmental Economics and Management* 42: 183–206.

Laxminarayan, R., and M.L. Weitzman. 2002. On the implications of endogenous resistance to medications. *Journal of Health Economics* 21: 709–718.

Malani, A. 2006. Identifying placebo effects with data from clinical trials. *Journal of Political Economy* 114: 236–256.

Mechoulan, S. 2004. HIV testing, a Trojan horse? *Topics in Economic Analysis and Policy,* 4(1), article 18.

Oster, E. 2006. HIV and sexual behavior change: Why not Africa? Working paper, Graduate School of Business, University of Chicago.

Philipson, T. 1995. The welfare loss of disease and the theory of taxation. *Journal of Health Economics* 14: 387–396.

Philipson, T. 2000. Economic epidemiology and infectious disease. In *Handbook of Health Economics*, ed. T. Culyer and J. Newhouse. Amsterdam: North-Holland.

Philipson, T., and J. DeSimone. 1997. Experiments and subject sampling. *Biometrika* 84: 618–632.

Philipson, T., and L. Hedges. 1998. Subject evaluation in social experiments. *Econometrica* 66: 381–409.

Philipson, T., S. Mechoulan, and A.B.. Jena. 2006. IP and external consumption effects: Generalizations from health care markets. Working Paper No. 11930. Cambridge, MA: NBER.

Philipson, T., and R.A. Posner. 1993. *Private choices and public health, an economic interpretation of the AIDS epidemic*. Cambridge, MA: Harvard University Press.

Philipson, T., and R.A. Posner. 2003. The long run growth of obesity as a function of technological change. *Perspectives in Biology and Medicine* 46(3): 87–108.

Rashad, I., and M. Grossman. 2004. The economics of obesity. *Public Interest* 156: 104–112.

Economic Freedom

Alan Peacock

Economic freedom describes a particular condition in which the individual finds himself as a result of certain characteristics in his economic environment. Taking a simple formulation of decision-making in which it is assumed that the individual maximizes his satisfaction both as a consumer of private and government goods and services and as a supplier of factor services, his position may be depicted as follows:

$$\text{Max } U^i = U^i\left(x^i, q_k, a^i\right) \tag{1}$$

Subject to

$$p_k^c \cdot x^i + T_k^i = Y^i = \varphi\left(a^i\right) = p_k^a \cdot a^i \tag{2}$$

where x_i is a vector of 'private goods', q_k is a vector of goods supplied by government, a^i is a vector of factor inputs, p_k^c is a vector of product prices for private goods, p_k^a is a vector of factor prices, T_k^i is net tax liability of individual ^{i} (tax obligation *less* transfers), Y^i is personal income before tax of individual ^{i} and subscript k denotes an exogenously determined variable.

Assuming the budget constraint (2) is exactly satisfied, the individual maximizes his satisfaction solving for the vector of private-goods consumption in terms of their prices, disposable income and predetermined levels of public goods available for consumption, where goods prices and factor prices, quantities of factor inputs and tax liabilities are either known or predicted by the individual.

Economic freedom requires that the various terms in the budget constraint reflect the absence of 'preference or restraint' (Adam Smith) on the individual. Therefore p_k^c is a vector of product prices which result from the operation of competitive market forces with the individual being free to choose between alternatives. Similarly, p_k^a must be characterized by competition in the factor market with the individual being 'free to bring both his industry and capital into competition with those of any other man or order or men' (Adam Smith). There is less certainty concerning the constraints placed on T_k and q_k. Some writers would argue that economic freedom requires a pre-established limit on the values of T_k and q_k either expressed or implied in a country's constitution (Nozick's 'minimal state'; see Nozick 1974). Others would argue that within a system of democratic government it should be possible to devise voting systems through which individuals express their preferences for values of T_k and q_k which simulate if they do not replicate the competitive market in the private sector (see Buchanan 1975). All agree, however, that economic freedom is not compatible with large values of T and q in relation to values of x^i, mainly because a large public sector increases the monopoly power of public servants both as suppliers of public goods and factor services to produce them and encourages the growth of private monopolies as a defence against public monopsony buying.

Economic Freedom and Libertarian Philosophy

There are features of this attempt at a 'technical' definition which may be called in question and which must be considered later, but it will be recognizable to those economists who have elevated economic freedom to an important goal in its own right and have claimed that it is the most important means for ensuring that the economy develops at the right 'tempo'. Discussion of the usefulness of the concept of economic freedom, therefore, centres in these two libertarian propositions.

The first proposition is contained in a striking passage in Book III of his *Essay on Liberty*: J.S. Mill wrote:

> He who lets the world, or his own portion of it, choose his plan of life for him, has no need of any other faculty than the ape-like one of imitation. He who chooses to plan for himself, employs all his faculties. He must use observation to see, reasoning

> and judgment to foresee, activity to gain materials for decision, discrimination to decide and, when he has decided, firmness and self-control to hold to his deliberate decision. . . It is possible that he might be guided on some good path, and kept out of harm's way, without any of these things. But what will be his comparative worth as a human being? It really is of importance, not only what men do, but also what manner of men they are that do it. (Mill 1859)

The passage captures the essence of the libertarian view of the good society, clearly implying that it requires that individuals should accept the necessity for choosing and for recognizing their responsibility for making choices. It must simultaneously require that, to develop the capacity for choosing, individuals must have the widest possible freedom of choice in the acquisition and disposal of resources. Two further conclusions follow.

The only restriction on economic freedom experienced by the individual should be when such freedom harms others.

The individual is not accountable to society for his actions and this, together with the different and changing preferences of individuals, makes libertarians distance themselves from attempts to establish a 'social welfare function' (cf. Rowley and Peacock 1975).

The second proposition maintains that economic freedom brings the added bonus of promoting the economic welfare of both the individual and of society. Economic freedom encourages the individual to 'better his condition' (Smith 1776) by exploiting opportunities for specialization and gains from trade which will be fully realized through the spontaneous emergence of markets. Not only is economic freedom regarded as the only material condition compatible with human dignity but it is also a necessary condition for the economic growth of the economy and for its adjustment to the changing preference structures of its members in response to market forces. The market is a 'discovery process' (Hayek 1979) in which participants adjust to change giving rise to the notion of the 'invisible hand' which coordinates human economic actions automatically without recourse to government intervention. *Pace* Hahn (1982) and others, libertarians do not attach importance to a general equilibrium solution, attained by the operation of competitive market forces (cf. Barry 1985). Indeed, though some exceptions will be noted below, it is claimed by supporters of the doctrine of economic freedom that disturbance of the natural process of exchange by government intervention assumes knowledge of the intricacies of the economy which is vouchsafed to no one, but there is no guarantee that officials, who maximize their private interests like everyone else, would be willing to maximize some social optimum even if they knew how to do so.

It was clearly recognized, by Hume and Smith for example, that for markets to work efficiently there must be a well-defined system of property rights and that costs of contracting between individuals in order to benefit from gains-from-trade would need to be minimized. The promotion of market efficiency was therefore bound to require some government intervention. No specialization or gains-from-trade would take place in a society in which there was no machinery for settling disputes and for preserving law and order. Acceptance of coercive intervention, however, requires that the 'rule of law' prevails. The law must be prospective and never retrospective in its operation, the law must be known and, as far as possible, certain, and the law must apply with equal force to all individuals without exception or discrimination. The state could also have a role in reducing the costs of contracting both by the removal of barriers to trade and to factor mobility and by the positive encouragement to the reduction in the costs of transport. In this latter respect Adam Smith supported reduction in the 'expense of carriage' by state financing of road building and supervision of financial methods to promote road maintenance and improvement.

At no stage therefore in the development of the doctrine of economic freedom, as understood by economists, was it regarded as synonymous with 'laissez-faire'. At the same time, the role of the state in respect of the promotion of economic freedom was and has remained strictly limited in libertarian thinking. Indeed, some modern libertarians devote much discussion to the possibilities of 'privatizing' even such traditional functions of the state as the maintenance of law and order.

Some Problems Raised by the Concept of Economic Freedom

The most obvious question posed to libertarians by those who are sceptical of their position is that the system of economic freedom is silent on the question of the distribution of property rights. In terms of our simple model, what principle should determine the values of $Y^1, \ldots, Y^i, \ldots, Y^n$ which, when aggregated, would describe some initial distribution of income as measured, say, by the shape of the Lorenz Curve? What reason have we for supposing that the 'optimal' distribution of income would emerge from the process of economic exchange between individuals?

The answer to this question does not find libertarians speaking with one voice. The problem is not one of principle, for the ultimate test to them is how far any government intervention represents a restriction of freedom. The problem is one of interpretation. It would be difficult today to find libertarians who would object to government intervention designed to assure protection to those who are severely deprived. Thus Hayek has argued that so long as 'a uniform minimum income is provided outside the market to all those who, for any reason, are unable to earn in the market an adequate maintenance, this need not lead to a restriction of freedom, or conflict with the Rule of Law'. This still leaves room for much disagreement among libertarians as to the precise level of the minimum and how to decide on who is entitled to receive it. Some supporters of the libertarian position, including the present author, would go much further and argue, along with J.S. Mill, that concentrations of wealth sustained over lengthy time periods can endanger economic freedom, not to speak of political freedom, by the association of such concentrations with the concentration of power of wealthy individuals over the less fortunate.

If the concept of economic freedom cannot embrace some precise guidance about the extent to which economic exchanges should be interfered with, it certainly places limits on the form of that interference. Thus libertarians, to the extent that they accept the need for a state-guaranteed minimum standard of living, prefer the use of money transfers to individuals rather than the provision of social services below or at zero cost, that is to say the economic condition of individuals in receipt of state support should be reflected in reduction in T_k^i (whose value may have to be negative) rather than an increase in q_k. Thus it is argued that individuals then retain responsibility for the purchase of goods and services designed to promote their own welfare and that the power of the state over the individual by bureaucratic dictatorship of preferences and by the lack of incentives in the public sector to economize in resource use is circumscribed.

A more severe test for the practicality of libertarian measures, designed to permit some redistribution without increasing the power of the state, arises in the case of any attack on the concentration of wealth. Clearly, a system of inheritance taxation which results in the transfer of capital from the private to the public sector would not conform to libertarian thinking, not only because this would discourage private saving but also because it would build up the power of the state. A system of taxation would have to be devised which not only did not discourage accumulation of private capital but also simultaneously encouraged legators to disseminate capital in favour of those with little capital. It is a long time since libertarians have plucked up the courage to try to develop such a system, given that eminent public finance specialists have failed in their attempts to fulfil these requirements.

The second major question arises from the persistent objection of Marxists and other Socialist writers that the system of economic freedom, as depicted by the libertarians, fails to solve the problem of 'worker alienation'. It may be that the system of economic freedom can allow employees alone or in combination with others to influence the price of factor inputs (p_k^a) and the work/leisure combination (a^i), variables which play a crucial part in individual welfare. The fact remains that the system of property rights, which libertarians support, includes the individual ownership of capital and the use of capitalistic methods of production which imply an authority relationship between employer and worker. The hierarchical order at the place of work

seems at complete variance with the independence of economic action attributed to the individual by the supporters of economic freedom.

Reactions to this argument by libertarians are sometimes reminiscent of the Scots preacher who, on recognizing a theological difficulty in his sermon, recommended his congregation to look the difficulty squarely in the face and pass it by. However, even Socialist writers, notably the prominent Marxist Ota Sik (1974), have recognized that the alternative to market capitalism – collectivist production – does not solve the problem for it is not synonymous with democratization at the shop-floor level. In other words, the basis of alienation is technological and not institutional. Some libertarians, notably Mill, have made common cause with Socialists by arguing that alienation must not be taken to be an inevitable consequence of productive activity. Mill sought one solution in the encouragement of firms owned and managed by the labour force, but still subject to competition. Utopian Socialists have claimed that the only solution is to reject altogether the technology which imposes hierarchical relations in the first place. Both 'solutions' are still the subject of living debate in both the professional and political arena.

See Also

- ▶ Economic Harmony
- ▶ Self-interest
- ▶ Utilitarianism

Bibliography

Barry, N.P. 1985. In defense of the invisible hand. *The Cato Journal* 5(1): 133–148.

Buchanan, J.M. 1975. *The limits of liberty: Between Anarchy and Leviathan*. Chicago/London: University of Chicago Press.

Hahn, F. 1982. Reflections on the invisible hand. *Lloyds Bank Review* 144: 1–21.

Hayek, F.A. 1979. *Law, legislation and liberty*, vol. 3. London: Routledge & Kegan Paul.

Mill, J.S. 1859. *Essay on liberty*. Oxford: Oxford University Press, 1942.

Mill, J.S. 1871. *Principles of political economy*, Book IV, ch. VII. Toronto: University of Toronto Press, 1965.

Nozick, R. 1974. *Anarchy, state and utopia*. Oxford: Basil Blackwell.

Peacock, A. 1979. *The economic analysis of government and related themes*. Oxford: Martin Robertson, chs. 5 and 6.

Rowley, C., and A. Peacock. 1975. *Welfare economics: A liberal re-interpretation*. London: Martin Robertson.

Sik, O. 1974. The shortcomings of the Soviet economy as seen in Communist ideologies. *Government and Opposition* 9(3): 263–276.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, Book IV, ch. 9 and Book V. Ed. R.H. Campbell and A.J. Skinner. Oxford: Clarendon Press, 1976.

Economic Governance

Avinash K. Dixit

Abstract

Economic governance consists of the processes that support economic activity and economic transactions by protecting property rights, enforcing contracts, and taking collective action to provide appropriate physical and organizational infrastructure. These processes are carried out within institutions, formal and informal. The field of economic governance studies and compares the performance of different institutions under different conditions, the evolution of these institutions, and the transitions from one set of institutions to another.

Keywords

Accountability; Arbitration; Authoritarianism and economic development; Better Business Bureaus (USA); Coase, R.; Colonization; Commitment; Common law vs civil law; Common-pool resource management; Corporate governance; Corruption; Corruption control; Credit rating agencies; Democracy and economic development; Democracy: parliamentary vs presidential; Discriminating alignment hypothesis; Economic governance; Electoral rules; Ethnic trading networks; Evolution of institutions; Exploitation; Formal vs

informal (private) contract enforcement; For-profit private institutions; Governance; Growth and governance; Incentive compatibility; Inequality; Information verifiability; Institutions and transition; International contracts; Mafia; Market institutions; Monitoring mechanisms; Opportunism; Political stability; Prisoner's Dilemma; Property rights enforcement; Regulatory burden; Relation-based vs rule-based governance; Reputation mechanisms; Rule of law; Slavery; Social networks; Social norms; Subgame perfection; Title registration; Transaction cost economics; Trust; Voice; World Bank; World Trade Organization; Yakuza

JEL Classifications
D02; H11; K14; P26; P51

Formal and informal institutions arise and evolve to underpin economic activity and exchange by protecting property rights, enforcing contracts, and collectively providing physical and organizational infrastructure. The field of economic governance studies and compares these institutions: state politico-legal institutions, private ordering within the law (credible contracting, arbitration), for-profit governance (credit-rating agencies, organized crime), and social networks and norms. Private institutions can outperform the state's legal system in obtaining and interpreting relevant information, and imposing social sanctions on the violators of norms. But private institutions are often limited in size; as economic activity expands, a transition towards more formal institutions is usually observed.

Concepts and Taxonomies

The term 'governance' has exploded from obscurity to ubiquity in economics since the 1970s. A search of the EconLit database shows clear evidence of this explosion. In the relevant categories (title, keywords and abstracts), there are just five occurrences of the word from 1970 to 1979. The number jumps to 112 for the 1980s and 3,825 for the 1990s. Since 2000 to the time of this writing (December 2005), there are already 7,948.

The Oxford English Dictionary gives several definitions of the word 'governance': (*a*) the action or manner of governing; controlling, directing, or regulating influence; control, sway, mastery; the state of being governed; good order; (*b*) the office, function, or power of governing; authority or permission to govern; that which governs; (*c*) the manner in which something is governed or regulated; method of management, system of regulations; a rule of practice, a discipline; and (*d*) the conduct of life or business; mode of living; behaviour, demeanour; discreet or virtuous behaviour; wise self-command. These diverse meanings allow the word to be used (and sometimes misused) for almost any context of economic decision-making or policy.

Two areas of application merit special mention. One is *corporate governance*. This analyses the internal management of a corporation – organizational structure and the design of incentives for managers and workers – and the rules and procedures by which the corporation deals with its shareholders and other stakeholders.

The second is *economic governance*; Williamson (2005) expresses its theme as the 'study of good order and workable arrangements'. This includes the institutions and organizations that underpin economic transactions by protecting property rights, enforcing contracts, and organizing collective action to provide the infrastructure of rules, regulations, and information that are needed to lend feasibility or workability to the interactions among different economic actors, individual and corporate. Different economies at different times have used different institutions to perform these functions, with different degrees of success. The field of economic governance studies and compares these different institutions. It includes theoretical models and empirical and case studies of the performance of different institutions under different circumstances, of how they relate to each other, of how they evolve over time, and of whether and how transitions from one to another occur as the nature and scope of economic activity and its institutional requirements change.

Corporate governance and economic governance are connected because the boundary of a corporation is itself endogenous, determined by the same considerations of information and commitment costs that raise problems of internal organization as well as those of property and contract (Coase 1937). Specifically, the nature of transaction costs may make it more efficient to handle some problems of governance by merging the two parties, for example by vertical integration (Williamson 1975, 1995). But it is analytically convenient to separate the two. This article concerns economic governance. To avoid constant repetition, I will simply call it 'governance' here unless some explicit reference to corporate governance is relevant.

Governance was neglected by economists for a long time, perhaps because they expected the government to provide it efficiently. However, experience with less developed and reforming economies, and observations from economic history, have led economists to study non-governmental institutions of governance.

Governance is not a field per se; it is an organizing or encompassing concept that bears on issues in many fields, including institutions and organizational behaviour, economic development and growth, industrial organization, law and economics, political economy, comparative economic systems, and various subfields of these.

We can organize the subject by classifying institutions along different dimensions. As is usual with such taxonomies, these are conceptual categories to help organize our thinking and analysis. In reality, there are significant differences within each category and overlaps across categories.

The first dimension concerns the purpose of the institution. The categories are: (*a*) protection of property rights against theft by other individuals and usurpation by the state itself or its agents, (*b*) enforcement of voluntary contracts among individuals, and (*c*) provision of the physical and regulatory infrastructure to facilitate economic activity and the functioning of the first two categories of institutions. We might also consider a fourth category, namely, the deep institutions that are essential to avoid serious cleavages or alienation that threaten the cohesion of the society itself. But this has not been studied in this context so far.

The second dimension concerns the nature of the institution. The categories are: (*a*) the formal state institutions that enact and enforce the laws, including the legislature, police, judiciary and regulatory, agencies, (*b*) institutions of private ordering that function under the umbrella of state law, for example various forums for arbitration, (*c*) private for-profit institutions that provide information and enforcement, and (*d*) self-enforcement within social or ethnic groups and network. My discussion is organized in sections along this dimension.

A third dimension distinguishes institutions that arise and evolve organically from those that are designed purposively; self-enforcing groups are often organic while the first three categories in the second dimension usually require some measure of design. This matters for the evolution of institutions of governance (see Greif 2006, especially ch. 6; Williamson 2005 p. 1).

Formal Institutions of the State

There is broad agreement that the quality of institutions of governance significantly affects economic outcomes. The importance of protecting property rights, both from other individuals and from predation by the state itself, is generally recognized and documented (for example, De Soto 2000). But serious disputes about the precise measures of quality of institutions, and about many details of the causal mechanisms by which they affect economic outcomes, remain.

At the broadest level, the distinction is between democracy and authoritarianism, each of which comes in many different varieties. Democracy has many normative virtues, but its worth in governance is less clear. Barro (1999, p. 61) finds an inverse U-shaped relationship between economic growth and a continuous measure of democracy – 'more democracy raises growth when political freedoms are weak, but depresses growth when a moderate amount of freedom is already established' – but the fit is relatively poor. Persson (2005), using cross-sectional as well as

panel data, finds that the crude distinction between democratic and non-democratic forms of government is not enough. The precise form of democracy matters for policy design and economic outcomes: 'parliamentary, proportional, and permanent democracies seem to foster the adoption of more growth-promoting structural policies, whereas ... presidential, majoritarian, and temporary democracy do not' (Persson 2005, p. 22). However, Keefer (2004, p. 10), after surveying a wide-ranging literature on electoral rules and legislative organizations, concludes that they affect policies but are not a crucial determinant of success: 'electoral rules ... almost surely do not explain why some countries grow and others do not', and 'the mere fact that developing countries are more likely to have presidential forms of government is unlikely to be a key factor to explain slow development.'

Democracy can be important for governance because its reliance on rules and procedures provides citizens with protection against predation by the state or its agents. Indeed, the elite, which might otherwise prefer to rule unconstrained, may find it in its own interest to make a credible commitment not to steal from the population by creating and fostering democracy (Acemoglu 2003; Acemoglu and Robinson 2005). Greif et al. (1994) discuss how groups of traders (guilds) in late medieval Europe took collective action to counter rulers' incentives to violate their members' property rights.

Even in a democracy, agents of the state may pursue their private interests using corruption, complex regulations to extract rent, and favouritism. In fact, an emerging literature argues that economic growth, at least in its early stages, is better promoted under suitably authoritarian regimes. Glaeser et al. (2004) argue that less developed countries that achieve economic success do so by pursuing good policies, often under dictatorships, and only then do they democratize. While these conclusions are controversial, these authors' criticisms of the measures of institutions used in the research that argues for the primacy of institutions in general, and of democracy in particular, are telling. Giavazzi and Tabellini (2005) find a positive feedback between economic and political reform, but they also find that the sequence of reforms matters, and countries that implement economic liberalization first and then democratize do much better in most dimensions than those that follow the opposite route. In practice, of course, it is difficult to ensure *ex ante* that an authoritarian ruler will implement good governance.

Many different measures of institutional quality exist. World Bank researchers Kaufman et al. (2005, which contains citations to their earlier work) have constructed six: (*a*) Voice and Accountability – measuring political, civil and human rights; (*b*) Political Instability and Violence – measuring the likelihood of violent threats to, or changes in, government, including terrorism; (*c*) Government Effectiveness – measuring the competence of the bureaucracy and the quality of public service delivery; (*d*) Regulatory Burden – measuring the incidence of market-unfriendly policies; (*e*) Rule of Law – measuring the quality of contract enforcement, the police, and the courts, as well as the likelihood of crime and violence; and (*f*) Control of Corruption – measuring the exercise of public power for private gain, including both petty and grand corruption and state capture. Of these, (*e*), (*f*) and also (*b*) concern the most basic institutions for protection of property rights and enforcement of contracts, (*a*) relates to governance because voice and accountability can reduce the severity of the agency problem between the citizens and the agencies of the state, and (*c*) and (*d*) pertain to what I called provision of the infrastructure of governance. Conceptually they are a mixed bag; the quality of some of them can itself depend on the quality of other more basic ones, and some are closer to being measures of effects than of causes. Their method of construction relies on subjective perceptions, and is subject to error. But when used with caution, they have proved significant as explanatory variables in empirical studies of economic growth, and for observing changes in governance quality over time in specific countries. Corruption and regulatory burdens are major themes of the World Bank's research on governance in many countries (see World Bank Institute, website).

Empirical estimations of the level or growth of GDP on various measures of institutional quality confront many conceptual and econometric problems. Researchers have tackled the issue of reverse causation by using various instruments, such as the nationality of colonizers (Hall and Jones 1999), mortality among colonizers (Acemoglu et al. 2001), and whether a colony had rich mineral resources or climatic and soil conditions conducive to plantation agriculture and a large or dense native population, or was sparsely populated and poor in the 1500 s (Engerman and Sokoloff 2002; Acemoglu et al. 2002). The general idea is that in the former circumstances the European colonizers established institutions of slavery and inequality to facilitate the exploitation of labour on a large scale, whereas in the latter conditions, where the colonizers had to exert their own effort, their institutions provided the correct incentives and became conducive to longer-term economic success. The debate on the factual and econometric validity and the economic interpretation of these findings is fierce and continuing; Hoff (2003) surveys and discusses this literature in detail.

La Porta et al. (1998, 1999) contrast different legal traditions for protecting the rights of small shareholders. If such protection is poor, that will inhibit the flows of capital to its most efficient uses. They find that systems based on common law are better in this regard than those based on civil law. But Rajan and Zingales (2003) and Lamoreaux and Rosenthal (2005) argue that in practice there was little difference between the systems during critical periods of industrialization.

These debates are sure to continue, and this section will get out of date very quickly.

At the international level, formal governance works through bodies like the World Trade Organization. Their members are sovereign countries; therefore their procedures must be subject to self-enforcement in repeated interactions, whether through bilateral or multilateral sanctions. These institutions are therefore basically similar to the social networks discussed below. See Maggi (1999) and Bagwell and Staiger (2003) for detailed analyses.

Private Institutions

The policing functions for property right protection supplied by the state are often supplemented by private security systems that serve specific clients and purposes – firms employ or hire security personnel, gated communities and neighbourhoods have private (hired or volunteer) patrols. These generally merely supplement the functions of the police for their specific context and work cooperatively with the police, but the two may clash if the private security system goes beyond its permissible functions.

Private institutions of contract enforcement similarly coexist with formal law, and become essential when the latter is weak or nonexistent. Explicit or implicit private contractual arrangements are also important for assignment of property rights as a part of Coasean contracting for efficient outcomes. Therefore, analyses of private institutions often focus on the governance of contracts.

The basic problem of contract enforcement is control of opportunism. If one or both parties have to make transaction-specific investments, the other can attempt to secure a greater part of the benefit by reneging or demanding renegotiation. The prospect of this can jeopardize the potentially mutually beneficial deal in the first place. Williamson (1975, 1995) pioneered the analysis of this issue under the title of transaction cost economics.

Information constitutes a major source of advantage for private ordering over formal law. Enforcement of a contract in a court requires offering proof of misconduct by the other party in the dispute; the relevant information must be verifiable to outsiders. Therefore, formal contracts can stipulate actions by the parties conditional only on verifiable information. Other or more detailed information may be observable to the parties themselves, or can be inferred by specialist insiders to the industry, but cannot be verified to non-specialist judges or juries of the state's legal system, or can be verified only at excessive cost.

The informational advantage of private ordering may be offset by a disadvantage in enforcement. Informal arrangements must be made to

overcome each participant's temptation to behave opportunistically at the others' expense. Different methods of this kind underlie the various institutions of informal governance, and achieve different degrees of success. Some are able to exert coercion for immediate punishment of misbehaviour. Others create long-run costs, typically in the form of exclusion from future participation or worse future opportunities, to offset the short-run advantages of opportunism. This is the standard theory of self-enforcing cooperation in repeated Prisoner's Dilemmas. The following sections discuss some of these alternatives.

Private Ordering with Formal Law in the Background

Perhaps the most remarkable thing about formal legal institutions and mechanisms for the enforcement of commercial contracts is how rarely they are actually used. Business transactions often do have underlying formal contracts, but when disputes arise recourse to the law is often the last resort. Other private alternatives are tried first; these include bilateral negotiation, arbitration by industry experts, and so on. Filing a suit in a formal court of law often signals the end of a business relationship. Most actual practice in business contracting is therefore better characterized as 'private ordering under the shadow of the law' (Macaulay 1963; Williamson 1995, pp. 95–100, 121–2).

If one of the parties to an ongoing informal relationship behaves opportunistically, the most common alternative is to fall back on a formal contract based on verifiable contingencies alone. Suppose an outcome based on a tacit understanding of what each party should do in any one exchange (including good-faith negotiation to adapt to changing circumstances) yields both of them higher payoffs than does a formal contract. Consider the implicit arrangement where, if one party deviates from the agreed course of action to its own advantage and to the detriment of the other, their future exchanges will be governed by the formal contract. This yields a subgame-perfect (credible) equilibrium of the repeated game if each party's one-time gain from opportunism does not exceed the capitalized value of the future difference of payoffs between the tacit and the formal contracts. Williamson (2005, p. 2) expresses this well: 'continuity can be put in jeopardy by defecting from the spirit of cooperation and reverting to the letter.'

When such relationship-based implicit contracting prevails, partial improvement in the formal system can worsen the outcome, due to a problem of the second-best. The partial improvement raises the payoffs the two parties could get from the fallback formal contract. This in turn reduces the future cost of a current deviation from the implicit contract or spirit of cooperation. It tightens the incentive-compatibility constraints, and therefore worsens what can be achieved by relational contracting (Baker et al. 1994; Dixit 2004, ch. 2).

Arbitration comes in two prominent forms. One is industry-specific, based on expert knowledge of insiders. More information is verifiable in such settings; therefore richer contracts specifying actions for more detailed contingencies become feasible. In many industries there is a large common-knowledge basis of custom and practice, which may even make it unnecessary to write down a contingent contract in great detail. Arbitration can also provide an opportunity for the parties to communicate and renegotiate adaptations to new circumstances. Formal legal systems often recognize these advantages of expert arbitration, and courts stand ready to enforce the decisions of arbitrators if the losing party tries to evade the sanctions. However, industry arbitrators often have severe sanctions at their own disposal; they can essentially drive the miscreant out of business, and even ostracize him or her from the social group of that business community. Examples of arbitration institutions include Bernstein's (1992) classic study of the diamond industry. For further discussion and modelling, see Dixit (2004, ch. 2) and Williamson (2005, p. 14).

The other prominent forums of arbitration deal with international contracts (Dezalay and Garth 1996; Mattli 2001). There are several of these, specializing in different legal traditions. They lack direct power to enforce their decisions, but

are backed by treaties that ensure enforcement by national courts. These forums do not have industry-specific knowledge, their processes can be slow and costly, and their decisions can be somewhat arbitrary. But parties in transnational transactions may prefer them to either country's courts, suspecting that these will be biased in favour of their own nationals.

For-Profit Private Institutions

If the state is unwilling to protect certain kinds of property or enforce certain kinds of contracts (for example in illegal activities), or is unable to do so (for example in weak and failing states), or is itself predatory, then private institutions can emerge to perform these functions for a profit. Organized crime often fills the niches uncovered by the state. Gambetta (1993), Bandiera (2003) and others argue that the Mafia emerged in just such a situation to fill the vacuum of protection in late 19th-century Sicily. Landowners began to hire guards of former feudal lords, and even the toughest among bandits, to protect their property. Gambetta describes how the Mafia's role expanded to providing contract enforcement in illegal or grey markets. Similarly, the Japanese Yakuza was instrumental in organizing markets at the end of the Second World War in August and September 1945 when the Japanese state had collapsed (Dower 1999, pp. 140–8), and mafias grew in Russia after the collapse of the Soviet regime (Varese 2001).

Gambetta (1993, p. 19) argues that this 'business of protection' is the core business of the Mafia. It may engage in other activities using in-house protection, but that is just downstream vertical integration – the opposite of upstream integration where an ordinary business firm has its in-house security department. A transaction-cost analysis of the internal organization of mafias, and of their vertical integration decisions, may provide an interesting link between economic governance and corporate governance. Another dimension in which the protection business can expand is extortion; although private protectors may be welcome when state protection has collapsed, 'protectors, once enlisted, invariably overstay their welcome' (Gambetta 1993, p. 198).

The Mafia can provide contract enforcement because, even though two traders may not have sufficiently frequent dealings with each other to achieve good outcomes in an ongoing bilateral relationship, each trader can be a regular customer of the enforcer. This converts multiple one-shot Prisoner's Dilemma games among the whole group of traders into several bilateral repeated games of each trader with the enforcer. The intermediary can provide information (keeping track of previous contract violations and informing a customer of the history of a potential trading partner) and/or actual punishment if a customer's trading partner violates their contract. The information role of the Mafioso is similar to that of credit rating agencies and Better Business Bureaus in the United States. Dixit (2004, ch. 4) constructs a model of such for-profit governance, and establishes the conditions for an equilibrium with for-profit private enforcement. These are lower bounds on the shares of the surplus that the customer and the Mafioso must have, so as to overcome the trader's temptation to cheat and the Mafioso' temptation to double-cross the customer. Milgrom et al. (1990) have a related and complementary model of private judges at medieval European trade fairs. They specify the game of each trade, and investigation in the event of cheating, in greater detail, but do not examine the issue of the judges' honesty.

Group Enforcement Through Social Networks and Norms

Any institution of contract enforcement must solve three key problems: (*a*) detection of opportunistic deviations from the contractually stipulated behaviour, (*b*) preservation and dissemination of information about the histories of the participants' behaviour, and (*c*) inflicting appropriate punishments to reduce future payoffs of any deviators. The first is often constrained by the available technology of monitoring, although institutions and regulations such as reporting

requirements and auditing can improve the technology. The second and third problems are best resolved in bilateral ongoing relationships: each party has a natural incentive to detect and remember the other's cheating, and can punish the other by breaking off the relationship. However, governance is often needed in groups each of whose members interacts frequently with someone else in the group, but not necessarily bilaterally with the same person every time. Now remembering and transmitting information about your current partner's behaviour to others, and refusing a potentially beneficial deal because the counterparty has cheated someone else in the past, are privately costly activities and therefore require their own governance mechanisms.

Formal state institutions of governance can solve these problems by fiat; the legal system compels the whole group of traders to commit to good behaviour by subjecting themselves to detection and punishment if they cheat. A third-party supplier of information or enforcement serves similar functions. In the case of a Mafia enforcer, anyone who trades with a customer of the Mafioso subjects himself to the grim punishment if he cheats. In the case of a Better Business Bureau, a firm that joins the organization thereby gives hostage to its own good behaviour: if it misbehaves it will get a poor rating or blacklisting. Transactions vary in their characteristics; therefore we should expect the effectiveness of such reputation mechanisms to vary also, and should not expect universal success from any one.

An institution of social networks and norms can solve the problems of information and punishment in a decentralized manner. Each participant can transmit information about his or her current trading partner's behaviour to others in the group to whom he or she is linked. And each can play his or her assigned part in punishment, typically by refusing to trade, if he or she gets matched with a potential partner who is known to have misbehaved in past dealings with others in the group. Incentives to transmit information or refuse potentially good trades can be established by a norm that regards refusal to do so as itself a punishable offence, as in Abreu's (1986) penal codes for repeated games; see Calvert (1995a, b). Extrinsic incentives may even be unnecessary if people have sufficiently strong natural instincts to punish social cheaters, as found by Fehr and Gächter (2000).

Numerous empirical and case studies of governance based on social relations have been conducted; space constraints allow mention of only a few. Greif's (1993) historical analysis of Maghribi traders' system of communication and collective punishment is well known. So is Ostrom's (1990) synthesis of the evidence on common-pool resource management; she emphasizes the importance of local knowledge and communication, of appropriately designed (generally graduated) punishments, and of incentives for individuals to perform their assigned roles and actions in the system. Fafchamps (2004) studies and compares many different market institutions in Africa; his work highlights the importance of designing systems appropriate to the conditions of each country or group. Ensminger (1992) describes a similarly rich complex of arrangements for trade and employment relationships among the Orma tribe of Kenya, and examines how formal institutions of property right enforcement including title registration can interact dysfunctionally with traditional arrangements based on family and tribal connections. Johnson et al. (2002) present and analyse findings from survey research in former socialist economies. Of particular interest are the links between evolving formal and informal governance. Even without a backup of courts, trust in bilateral relationships can build quickly in response to good experiences. New or transient customers are more likely to be offered credit if courts work better, but the effectiveness of courts becomes largely irrelevant for the functioning of established relationships. Casella and Rauch (2002) study the role of ethnic networks in international trade.

Li (2003) points out a key difference between the costs of operating such a system and those of formal governance. A relation-based system of networks and norms has low fixed costs, but high and rising marginal costs. Trading on a small scale naturally starts among the most closely connected people who have sufficiently good

communication and common understanding to sustain honesty. No fixed costs need be incurred to establish any formal rules or mechanisms of enforcement. But as trade expands, potential partners added at the margin are almost by definition less well-connected, making it harder to communicate information with them and to ensure their participation in any punishments. By contrast, formal or rule-based governance has high fixed costs of setting up the legal system and the information mechanism, but once these are incurred, marginal costs of dealing with strangers are low. Therefore, relation-based governance is better for small groups and rule-based governance better for large groups. Greif's (1994) comparison between the relation-based system of Maghribi traders and the formal institutions of Genoese traders supports this theory. Dixit (2004, ch. 3) constructs a formal model that compares relation-based and rule-based systems. This characterizes the maximum size of a self-enforcing group, and finds that, when the group exceeds this critical size, the maximum scope of sustainable honesty shrinks absolutely. The intuition is as follows. At the critical size, each trader is indifferent between honesty and cheating when dealing with the most distant person. When more traders are added, this weakens the communication between the previously marginal person and other almost equally distant ones, tipping the balance toward cheating.

Kranton (1996) models individuals who can either choose bilateral long-lived self-enforcing trading relationships or search for one-time trading partners in an anonymous market with external enforcement. The market thus provides the outside opportunity in the repeated game of bilateral trade. If more people trade in the anonymous market, it becomes thicker and offers better prospects for successful search. Then parties in bilateral relationships have better outside opportunities, which makes it harder to sustain tacit cooperation there, further increasing the relative attraction of the market. Therefore the system can have multiple equilibria – no one uses the market because no one else uses it, or everyone uses the market because everyone else does – and can get locked into a Pareto-inferior equilibrium.

Evolution and Transformation of Governance Institutions

A persistent theme in this survey has been that different governance institutions are optimal for different societies, for different kinds of economic activity, and at different times. Changes in underlying technologies of production, exchange and communication change the relative merits of different methods of governance. As the volume and scope of trade expand, formal institutions generally become superior to informal ones, but informal ones serve useful roles under the shadow of formal ones even in the most advanced economies and sectors. All this raises the question of whether we should expect institutions to adapt and evolve optimally.

Williamson's famous 'discriminating alignment hypothesis' says that transactions, with their different attributes, align with institutions, with their different costs and competencies; see his recent exposition (2005, p. 6). This gives ground for optimism for synergistic evolution of the need for governance and the institutions that supply it. Others are less sanguine. North (1990) and others argue that institutional change is subject to long delays due to resistance by organized interests favouring the status quo, problems of coordinating collective action to bring about a discrete change in equilibrium, and so on. Dixit (2004, pp. 79–85) discusses some of these problems for transition from relation-based to rule-based contract enforcement. Eggertson (2005) gives a dramatic example of how institutions restricting fishing and requiring costly mutual insurance persisted in Iceland for centuries after they had become obstacles to good economic performance.

I believe that a balanced approach is needed, recognizing the tendency towards synergistic alignment but also the obstacles to its realization. The net outcome will depend on many specifics of each context. Understanding and predicting the process requires a combination of approaches: case-based and analytical, inductive and deductive. Greif (2006) discusses, develops and applies such methodologies using historical studies of trade in medieval Europe.

See Also

- Cooperation
- Corporations
- Growth and Institutions
- Hold-Up Problem
- Law, Economic Analysis of
- Law, Public Enforcement of
- Market Institutions
- Property Rights
- Social Norms
- Spontaneous Order
- Transition and Institutions

Acknowledgments I thank Tore Ellingsen, Diego Gambetta, Karla Hoff, Eva Meyersson-Milgrom, Dani Rodrik, Oliver Williamson, and the editors for comments on previous drafts, and the National Science Foundation for research support.

Bibliography

Abreu, D. 1986. Extremal equilibria of oligopolistic supergames. *Journal of Economic Theory* 39: 191–225.

Acemoglu, D. 2003. Why not a political Coase Theorem? Social conflict, commitment and politics. *Journal of Comparative Economics* 31: 620–652.

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.

Acemoglu, D., and J. Robinson. 2005. *Economic origins of dictatorship and democracy*. New York: Cambridge University Press.

Bagwell, K., and R. Staiger. 2003. *The economics of the world trading system*. Cambridge, MA: MIT Press.

Baker, G., R. Gibbons, and K. Murphy. 1994. Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics* 109: 1125–1156.

Bandiera, O. 2003. Land reform, the market for protection and the origins of the Sicilian Mafia: Theory and evidence. *Journal of Law, Economics, and Organization* 19: 218–244.

Barro, R. 1999. *Determinants of economic growth: A cross-country empirical study*. Cambridge, MA: MIT Press.

Bernstein, L. 1992. Opting out of the legal system: Extralegal contractual relations in the diamond industry. *The Journal of Legal Studies* 21: 115–157.

Calvert, R. 1995a. The rational choice theory of social institutions: Cooperation, communication, and coordination. In *Modern political economy: Old topics, new directions*, ed. J. Banks and E. Hanushek. Cambridge: Cambridge University Press.

Calvert, R. 1995b. Rational actors, equilibrium, and social institutions. In *Explaining social institutions*, ed. J. Knight and I. Sened. Ann Arbor: University of Michigan Press.

Casella, A., and J. Rauch. 2002. Anonymous market and group ties in international trade. *Journal of International Economics* 58: 19–47.

Coase, R. 1937. The nature of the firm. *Economica* 4: 386–406.

De Soto, H. 2000. *Mystery of capital: Why capitalism triumphs in the west and fails everywhere else*. New York: Basic Books.

Dezalay, Y., and B. Garth. 1996. *Dealing in virtue: International commercial arbitration and the construction of a transnational order*. Chicago/London: University of Chicago Press.

Dixit, A. 2004. *Lawlessness and economics: Alternative modes of governance*. Princeton: Princeton University Press.

Dower, J. 1999. *Embracing defeat: Japan in the wake of world war II*. New York: W.W. Norton.

Eggertson, T. 2005. *Imperfect institutions: Possibilities and limits of reform*. Ann Arbor: University of Michigan Press.

Engerman, S., and K. Sokoloff. 2002. Factor endowments, inequality, and paths of development among New World economies. *Economia* 3: 41–109.

Ensminger, J. 1992. *Making a market: The institutional transformation of an African society*. New York: Cambridge University Press.

Fafchamps, M. 2004. *Market institutions in Sub-Saharan Africa: Theory and evidence*. Cambridge, MA: MIT Press.

Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980–994.

Gambetta, D. 1993. *The Sicilian Mafia: The business of private protection*. Cambridge, MA: Harvard University Press.

Giavazzi, F., and G. Tabellini. 2005. Economic and political liberalizations. *Journal of Monetary Economics* 57: 1297–1330.

Glaeser, E., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2004. Do institutions cause growth? *Journal of Economic Growth* 9: 271–303.

Greif, A. 1993. Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review* 83: 525–548.

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. 2006. *Institutions and the path to the modern economy: Lessons from medieval trade*. New York: Cambridge University Press.

Greif, A., P. Milgrom, and B. Weingast. 1994. Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy* 102: 745–776.
Hall, R., and C. Jones. 1999. Why do some countries produce so much more output than others? *Quarterly Journal of Economics* 114: 83–116.
Hoff, K. 2003. Paths of institutional development: A view from economic history. *World Bank Research Observer* 18: 205–226.
Johnson, S., J. McMillan, and C. Woodruff. 2002. Courts and relational contracts. *Journal of Law, Economics, and Organization* 18: 221–277.
Kaufman, D., A. Kraay, and M. Mastruzzi. 2005. *Governance matters IV: Updated governance indicators 1996–2004*. Washington, DC: World Bank research paper.
Online. Available at http://www.worldbank.org/wbi/governance/pubs/govmatters4.html. Accessed 20 Apr 2006.
Keefer, P. 2004. What does political economy tell us about economic development – And vice versa? *Annual Review of Political Science* 7: 247–272.
Kranton, R. 1996. Reciprocal exchange: A self–sustaining system. *American Economic Review* 86: 830–851.
La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 1998. Law and finance. *Journal of Political Economy* 106: 1113–1155.
La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 1999. The quality of government. *Journal of Law, Economics, and Organization* 15: 222–279.
Lamoreaux, N., and J.-L. Rosenthal. 2005. Legal regime and contractual flexibility: A comparison of business's organizational choices in France and the United States during the era of industrialization. *American Law and Economics Review* 7: 28–61.
Li, J.S. 2003. Relation-based versus rule-based governance: An explanation of the East Asian miracle and Asian crisis. *Review of International Economics* 11: 651–673.
Macaulay, S. 1963. Non-contractual relationships in business: A preliminary study. *American Sociological Review* 28: 55–70.
Maggi, G. 1999. The role of multilateral institutions in international trade cooperation. *American Economic Review* 89: 190–214.
Mattli, W. 2001. Private justice in a global economy: From litigation to arbitration. *International Organization* 55: 919–947.
Milgrom, P., D. North, and B. Weingast. 1990. The role of institutions in the revival of trade: The law merchant, private judges, and the Champagne fairs. *Economics and Politics* 2: 1–23.
North, D. 1990. *Institutions, institutional change, and economic performance*. Cambridge: Cambridge University Press.
Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge/New York: Cambridge University Press.
Persson, T. 2005. *Forms of democracy, policy, and economic development*, Working Paper No. 11171. Cambridge, MA: NBER.
Rajan, R., and L. Zingales. 2003. The great reversals: The politics of financial development in the twentieth century. *Journal of Financial Economics* 69: 5–50.
Varese, F. 2001. *The Russian Mafia: Private protection in a new market economy*. Oxford: Oxford University Press.
Williamson, O. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.
Williamson, O. 1995. *The mechanisms of governance*. New York: Oxford University Press.
Williamson, O. 2005. The economics of governance. *American Economic Review* 95: 1–18.
World Bank Institute. Governance and anti-corruption. Online. Available at http://www.worldbank.org/wbi/governance. Accessed 20 Apr 2006.

Economic Growth

Peter Howitt and David N. Weil

Abstract

Economic growth is the increase in a country's standard of living over time. Growth economists study how living standards differ across countries as well as across time. This article discusses some of the broad facts of economic growth and some of the main approaches to its study.

Keywords

Adjustment costs; Capital accumulation; Cobb–Douglas functions; Competition policy; Creative destruction; Cultural determinants of growth; Demographic dividend; Development accounting; Dutch disease; Economic growth; Endogenous growth; Export promotion; Fertility; General purpose technology; Geographical determinants of growth; Growth accounting; Health; Human capital; Import substitution; Increasing returns; Inequality (global); Innovation; Institutional determinants of growth; Learning by doing; Malthus, T. R.; Mortality; Natural capital; Neoclassical growth theory; Patents; Physical capital; Policy determinants of growth; Population growth; Production functions; Productivity growth; Research and development; Resource curse; Savings and

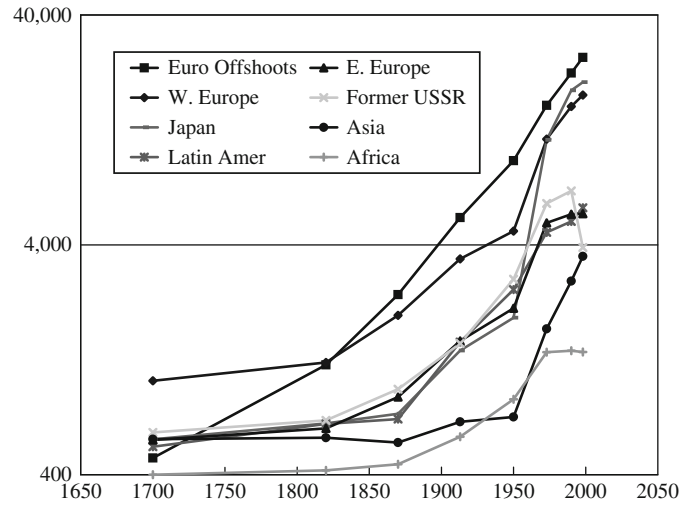
investment; Schooling; Technical change; Technical complementarities; Technology; Technology spillovers; Trade and growth

JEL Classifications
D4; D10

Economic growth is typically measured as the change in per capita gross domestic product (GDP). Sustained long-term economic growth at a positive rate is a fairly recent phenomenon in human history, most of it having occurred in the last 200 years. According to Maddison's (2001) estimates, per capita GDP in the world economy was no higher in the year 1000 than in the year 1, and only 53 per cent higher in 1820 than in 1000, implying an average annual growth rate of only one- nineteenth of one per cent over the latter 820-year period. Some time around 1820, the world growth rate started to rise, averaging just over one-half of one per cent per year from 1820 to 1870, and peaking during what Maddison calls the 'golden age', the period from 1950 to 1973, when it averaged 2.93 per cent per year. By 2000, world per capita GDP had risen to more than 8.5 times its 1820 value.

Growth has been uneven not only across time but also across countries. Since 1820, living standards in Western Europe and its offshoots in North America and the Antipodes have raced ahead of the rest of the world, with the exception of Japan, in what is often referred to as the 'Great Divergence'. As shown in Fig. 1 below, the proportional gap in per capita GDP between the richest group of countries and the poorest group (as classified by Maddison) grew from three in 1820 to 19 in 1998. Pritchett (1997) tells a similar story, estimating that the proportional gap between the richest and poorest countries grew more than fivefold from 1870 to 1990.

This widening of the cross-country income distribution seems to have slowed during the second half of the 20th century, at least among a large group of nations. Indeed, Fig. 1, which is drawn on a proportional scale, shows that with the acceleration of growth in Asia there has been a narrowing of the spread between the richest and the second poorest group since 1950. Evans (1996) shows a narrowing of the top end of the distribution (that is, among Organisation for Economic Cooperation and Development, OECD, countries) over the period. However, not all countries have taken part in this convergence process, as the gap between the leading countries as a whole and the very poorest countries has continued to widen. In Fig. 1 the gap between the Western Offshoots and Africa grew by a factor of 1.75 between 1950 and 1998. Likewise, the proportional income gap between Mayer-Foulkes's (2002) richest and poorest convergence groups grew by a factor of 2.6 between 1960 and 1995.



Economic Growth, Fig. 1 Per capita GDP, 1650–2000, 1990 international dollars (*Source:* Maddison (2001))

Jones (1997) argues that continuing divergence of the poorest countries from the rest of the world does not imply rising income inequality among the world's population, mainly because China and India, which contain about 40 per cent of that population, are rising rapidly from near the bottom of the distribution. Indeed, Sala-i- Martin (2006) shows, using data on within-country income distributions, that the cross-individual distribution of world income narrowed considerably between 1970 and 2000, even as the cross-country distribution continued to widen somewhat. But between-country inequality is still extremely important; in 1992 it explained 60 per cent of overall world inequality (Bourguignon and Morrison 2002). Another reason that growth economists are typically more concerned with the cross-country than the cross-individual distribution is that many of the determinants of economic growth vary across countries but not across individuals within countries.

The Production Function Approach

The main task of growth theory is to explain this variation of living standards across time and countries. One way to organize one's thinking about the sources of growth is in terms of an aggregate production function, which indicates how a country's output per worker y depends on the (per worker) stocks of physical, human, and natural capital, represented by the vector k, according to

$$y = f(k, A), \tag{1}$$

where A is a productivity parameter. Economic growth, as measured by the growth rate of y, depends therefore on the rate of capital accumulation and the rate of productivity growth. Similarly, countries can differ in their levels of GDP per capita either because of differences in capital or because of differences in productivity. Much recent work on the economics of growth has focused on trying to identify the relative contributions of these two fundamental factors to differences in growth rates or income levels among countries.

Modern growth theory started with the neoclassical model of Solow (1956) and Swan (1956), who showed that in the long run growth cannot be sustained by capital accumulation alone. In their formulation, the diminishing marginal product of capital (augmented by an Inada condition that makes the marginal product asymptote to zero as capital grows) will always terminate any temporary burst of growth in excess of the growth rate of labour-augmenting productivity. But this perspective has been challenged by more recent endogenous growth theory. In the *AK* theory of Frankel (1962) and Romer (1986), growth in productivity is functionally dependent on growth in capital, through learning by doing and technology spillovers, so that an increase in investment rates in physical capital can also sustain a permanent increase in productivity growth and hence in the rate of economic growth. In the innovation- based theory that followed *AK* theory, the Solow model has been combined with a Schumpeterian theory of productivity growth, in which capital accumulation is one of the factors that can lead to a permanently higher rate of productivity growth (Howitt and Aghion 1998).

Capital

Having introduced the production function in a general sense, we now examine the accumulation of different types of capital in more detail, and then turn to an assessment of the relative importance of factor accumulation and productivity in explaining income differences among countries and growth over time.

Physical Capital

Physical capital is made up of tools, machines, buildings, and infrastructure such as roads and ports. Its key characteristics are, first, that it is produced (via investment), and second that it is in turn used in producing output. Physical capital differs importantly from technology (which, as is discussed below, is also both produced and productive) in that physical capital is rival in its use: only a limited number of workers can use a single piece of physical capital at a time.

Differences in physical capital between rich are poor countries are very large. In the year 2000, for example, physical capital per worker was 148,091 dollars in the United States, 42,991 dollars in Mexico, and 6270 dollars in India. These large differences in physical capital are clearly contributors to income differences among countries in a proximate sense. That is, if the United States had India's level of capital it would be a poorer country. The magnitude of this proximate effect can be calculated by using the production function. For example, using a value for capital's share of national income of 1/3 (which is consistent with the findings of Gollin (2002) for a cross-section of countries), the ratio of capital per worker in the United States to that in India would by itself explain a ratio of income per capita in the two countries of 7.9 ($= (148{,}091/6270)^{1/3}$).

Differences in physical capital among countries can result from several factors. First, countries may differ in their levels of investment in physical capital relative to output. In an economy closed to external capital flows, the investment rate will equal the national saving rate. Saving rates can differ among countries because of differences in the security of property rights, due to the availability of a financial system to bring together savers and investors, because of government policies like budget deficits or pay-as-you-go old age pensions, differences in cultural attitudes towards present versus future consumption, or simply because deferring consumption to the future is a luxury that very poor people cannot afford.

A second factor that drives differences in investment rates among countries is the relative price of capital. The price of investment goods in relation to consumption goods is two to three times as high in poor countries as in rich countries. If one measures both output and investment at international prices, investment as a fraction of GDP is strongly correlated with GDP per capita (correlation of 0.50), and poor countries have on average between one half and one quarter of the investment rate of rich countries. When investment rates are expressed in domestic prices, the correlation between investment rates and GDP per capita falls to 0.05 (Hsieh and Klenow 2007).

But levels of capital can also differ among countries for reasons that have nothing to do with the rate of accumulation. Differences in productivity (the A term in Eq. 1) will produce different levels of capital even in countries with the same rates of physical capital investment. Similarly, differences in the accumulation of other factors of production will produce differences in the level of physical capital per worker.

Human Capital

Human capital refers to qualities such as education and health that allow a worker to produce more output and which themselves are the result of past investment. Like physical capital, human capital can earn an economic return for its owner. However, the two types of capital differ in several important respects. Most significantly, human capital is 'installed' in a person. This makes it very difficult for one person to own human capital that is used by someone else. Human capital investment is a significant expense. In the United States in the year 2000, spending by governments and families on education amounted to 6.2 per cent of GDP; forgone wages by students were of a similar magnitude.

Information on the productivity of human capital can be derived from comparing wages of workers with different levels of education. So called 'Mincer regressions' of log wage on years of education, controlling by various means for bias due to the endogeneity of schooling, yield estimated returns to schooling of about ten per cent per year. In the year 2000, the average schooling of workers in advanced countries was 9.8 years and among workers in developing countries 5.1 years. Applying a rate of return of ten per cent implies that the average worker in the advanced countries supplied 56 per cent more labour input because of this education difference. If labour's share in a Cobb–Douglas production function is two-thirds, this would imply that education differences would explain a factor of 1.35 difference in income between the advanced and developing countries, which is very small relative to the observed gap in income. Allowing for differences in school quality increases somewhat the income differences explained by human capital in the form of schooling.

A second form of human capital is health. The importance of health as an input into production can be estimated by looking at microeconomic data on how health affects individual wages. Health differences between rich and poor countries are large, and in wealthy countries worker health has improved significantly over the last 200 years (Fogel 1997). Weil (2007), using the adult survival rate as a proxy for worker health, estimates that eliminating gaps in worker health among countries would reduce the log variance of GDP per worker by 9.9 per cent.

Natural Capital

Natural capital is the value of a country's agricultural and pasture lands, forests and subsoil resources. Like physical and human capital, natural capital is an input into production of goods and services. Unlike other forms of capital, however, it is not itself produced.

Natural capital per worker and GDP per worker are positively correlated, but the link is much weaker than for the other measures of capital discussed above. The poor performance of many resource-rich countries has led many observers to identify a 'resource curse' by which the availability of natural capital undermines other forms of capital accumulation or reduces productivity. Among the suggested channels by which this happens are that resource booms lead countries to raise consumption to unsustainable levels, thus depressing saving and investment (Rodriguez and Sachs 1999); that exploitation of natural resources suppresses the development of a local manufacturing sector, which holds back growth because manufacturing is inherently more technologically dynamic than other parts of the economy (this is the so called Dutch disease); and that economic inefficiencies are associated with political competition or even civil war to appropriate the rents generated by natural resources.

Population and Economic Growth

Population affects the accumulation of all three forms of capital discussed above, and through them the level of output per worker. Rapid population growth dilutes the quantities of physical and human capital per worker, raising the rates of investment and school expenditure required to maintain output per worker. The interaction of natural capital with population growth is at the centre of the model of Malthus (1798). For a fixed stock of natural capital, higher population lowers output per capita. Combined with a positive feedback from the level of income to population growth, this resource constraint produces a stable steady state level of output per capita and, with technology fixed, a stable level of population as well. This Malthusian feedback is the explanation for the long period of nearly constant living standards that preceded the Industrial Revolution (Galor and Weil 2000). Because of resource-saving technological progress, as well as expansion of international trade, which allows countries to evade resource constraints, the interaction of population and natural capital is much less important today than in the past, with the exception of very poor countries that are reliant on subsistence agriculture.

In addition to its effect on the level of factors of production per worker, population also matters for economic growth because demographic change produces important changes in the age structure of the population. A reduction in fertility, for example, will produce a long period of reduced dependency, in which the ratio of children and the elderly, on the one hand, to working age adults, on the other, is temporarily below its sustainable steady state level. This is the so-called 'demographic dividend' (see population ageing).

In addition to these effects of population on the level of income per capital, there is also causality that runs from the economic to the demographic. Over the course of economic development, countries generally move through a demographic transition in which mortality rates fall first, followed by fertility rates. While the decline in mortality is easily explained as a consequence of higher income and technological progress, the decline in fertility is not fully understood. Among the factors thought to contribute to the decline in fertility are falling mortality, a shift along a quality–quantity trade-off due to rising returns to human capital, the rise of women's relative wages, the reduced importance

of children as a means of old age support, and improvements in the availability of contraception.

Growth Accounting and Development Accounting

The discussion above makes clear that stocks of different forms of capital are positively correlated with GDP per capita. Similarly, as countries grow, levels of capital per worker grow as well. It is natural to ask whether these variations in capital are sufficiently large to explain the matching variations in growth. The techniques of growth accounting (Solow 1957) and development accounting (Klenow and Rodriguez-Clare 1997; Hall and Jones 1999) attempt to give quantitative answers to this question. Using a parameterized production function and measures of the quantities of human and physical capital, one can back out relative levels of productivity among countries and rates of productivity growth within a country.

Caselli (2005) presents a review of development accounting along with his own thorough estimates. His finding is that if human and physical capital per worker were equalized across countries, the variance of log GDP per worker would fall by only 39 per cent. In other words, the majority of variation in income is due to differences in productivity, not factor accumulation. Differences in productivity growth, rather than differences in the growth of physical and human capital, are also the dominant determinants of differences in income growth rates among countries (Weil 2005, ch. 7; Klenow and Rodriguez-Clare 1997); differences in productivity levels among countries are striking. For example, comparing the countries at the 90th and 10th percentiles of the income distribution (which differ in income by a factor of 21), the former would produce seven times as much output as the latter with equal quantities of human and physical capital.

Productivity, Technology and Efficiency

Development accounting shows that productivity differences among countries are the dominant explanation for income differences. Similarly, differences in productivity growth are the most important explanation for differences in income growth rates among countries. And as a theoretical matter, the Solow model shows that as long as there are decreasing returns to capital per worker, productivity growth can be the only source of long-term growth. The question is: what explains these changes over time and differences in the level of productivity? Over the long term it is natural to associate productivity growth with technological change. However, especially as an explanation for differences in productivity at a given point in time, a second possibility is that productivity differences reflect differences not in technology, in the sense of inventions, blueprints, and so on, but rather differences in how economies are organized and use available technology and inputs. We label this second contributor to productivity as 'efficiency'.

Technology

Technology consists of the knowledge of how to transform basic inputs into final utility. This knowledge can be thought of as another form of capital, an intangible intellectual capital. What distinguishes technology from human or physical capital is its non-rival character. For example, the knowledge that a particular kind of corn will be immune to caterpillars, or the knowledge of how to produce a 3 GHz CPU for a portable computer, can be used any number of times by any number of people without diminishing anyone's ability to use it again. By contrast, if you drive a lorry for an hour, or if you employ the skills of a doctor for an hour, then that lorry or those skills are not available to anyone else during that hour.

Different growth theories have different approaches to modelling the accumulation of technology – that is, technical progress. According to neoclassical theory, for example, the relationship between technology and the economy is a one-way street, with all of the causation running from technology to the economy. It portrays technical progress as emanating from a scientific progress that operates outside the realm of economics, and thus takes the rate of technical progress as being given exogenously.

This neoclassical view has never been accepted universally. Specialists in economic history and the economics of technology have generally believed that technical progress comes in the form of new products, new techniques and new markets, which do not spring directly from the scientific laboratory; instead they come from discoveries made by private business enterprises, operating in competitive markets, and motivated by the search for profits. For example, the transistor, which underlies so much recent technological progress, was discovered by scientists working for the AT&T telephone company on the practical problem of how to improve the performance of switch boxes that were using vacuum tubes. Rosenberg (1981) describes many other examples of scientific and technological breakthroughs that originated in profit-oriented economic activity.

What kept this view of endogenous technology from entering the mainstream of economics until recently was the difficulty of incorporating increasing returns to scale into dynamic general equilibrium theory. Increasing returns arise once one considers technology as a kind of capital that can be accumulated, because of its non-rival nature; that is, the cost of developing a technology for producing a particular product is a fixed set-up cost, which does not have to be repeated when more of the product is produced. Once the technology has been developed then there should be at least constant returns to scale in the factors that use that technology, on the grounds that if you can do something once then you can do it twice. But this means that there are increasing returns in the broad set of factors that includes the technology itself. Increasing returns creates a problem because it generally implies that a competitive equilibrium will not exist, at least not without externalities.

These technical difficulties were overcome by the new 'endogenous growth theory' introduced by Romer (1986) and Lucas (1988), which incorporated techniques that had been developed for dealing with increasing returns in the theories of industrial organization and international trade. The first generation of endogenous growth theory to enter the mainstream was the 'AAK theory', according to which technological progress takes place as a result of externalities in learning to produce capital goods more efficiently. The second generation was the innovation-based theory of Romer (1990) and Aghion and Howitt (1992), which emphasizes the distinction between technological knowledge and other forms of capital, and analyses technological innovation as a separate activity from saving and schooling.

Historically, technical progress has engendered much social conflict, because it involves what Schumpeter (1942) called 'creative destruction'; that is, new technologies render old technologies obsolete. As a result, technical progress is a game with losers as well as winners. From the handloom weavers of early 19th century Britain to the former giants of mainframe computing in the late 20th century, many people's skills, capital equipment and technological knowledge have been devalued and their livelihoods imperilled by the same innovations that have created fortunes for others.

The destructive side of technical progress shows up most clearly during periods when a new 'general purpose technology' (GPT) is being introduced. A GPT is a basic enabling technology that is used in many sectors of the economy, such as the steam engine, the electric dynamo, the laser or the computer. As Lipsey et al. (2005) have emphasized, a GPT typically arrives only partially formed, creates technological complementarities and opens a window on new technological possibilities. Thus it is typically associated with a wave of new innovations. Moreover, the period in which the new GPT is diffusing through the economy is typically a period of rapid obsolescence, costly learning and wrenching adjustment. Greenwood and Yorukoglu (1997) argue that the productivity slowdown of the 1970s is attributable to the arrival of the computer, and Howitt (1998) argues that the rapid obsolescence generated by a new GPT can cause per capita income to fall for many years before eventually paying off in a much higher standard of living.

New technologies are often opposed by those who would lose from their introduction. Some of this opposition takes place within the economic sphere, where workers threaten action against

firms that adopt labour-saving technologies and firms try to pre-empt innovations by rivals. But much of it also takes place within the political sphere, where governments protect favoured firms from more technically advanced foreign competitors, and where people sometimes vote for politicians promising to preserve traditional ways of life by blocking the adoption of new technologies.

The leading industrial nations of the world spend large amounts on R&D for generating innovations. In the United States, for example, R&D expenditures constituted between 2.2 and 2.9 per cent of GDP every year from 1957 to 2004. But not much cutting-edge R&D takes place outside a small group of countries. In 1996, for example, 73 per cent of the world's R&D expenditure, as measured by UNESCO, was accounted for by just five countries (in decreasing order of R&D expenditure they are the United States, Japan, Germany, France and United Kingdom). In the majority of countries that undertake very little measured R&D, technology advances not so much by making frontier innovations as by implementing technologies that have already been developed elsewhere. But the process of implementation is not costless, because technologies tend to be context- dependent and technological knowledge tends to be tacit. So implementation requires an up-front investment to adapt the technology to a new environment (see, for example, Evenson and Westphal 1995). This investment plays the same role analytically in the implementing country as R&D does in the original innovating country.

Implementation is important in accounting for the patterns of cross-country convergence and divergence noted above. This is because a country in which firms are induced to spend on implementation have what Gerschenkron (1952) called an 'advantage of backwardness'. That is, the further they fall behind the world's technology frontier the faster they will grow with any given level of implementation expenditures, because the bigger is the improvement in productivity when they implement any given foreign technology. In the long run, as Howitt (2000) has shown, this force can cause all countries that engage in R&D or implementation to grow at the same rate, while countries in which firms are not induced to make such investments will stagnate. But technology transfer through implementation expenditures is no guarantee of convergence, because the technologies that are being developed in the rich R&D-performing countries are not necessarily appropriate for conditions in poor implementing countries (Basu and Weil 1998; Acemoglu and Zilibotti 2001) and because financial constraints may prevent poor countries from spending at a level needed to keep pace with the frontier (Aghion et al. 2005).

Efficiency

The efficiency with which a technology is used is not likely to play a major role in accounting for long-run growth rates, because there is a finite limit to how high you can raise living standards simply by using the same technologies more efficiently. But there is good reason to believe that differences in efficiency account for much of the cross-country variation in the level of productivity.

Inefficiencies take several different forms. Economic resources are sometimes allocated to unproductive uses, or even unused, as when union featherbedding agreements kick in. Resources can be misallocated as the result of taxes, subsidies and imperfect competition, all of which create discrepancies between marginal rates of substitution. Technologies can be blocked by those who would lose from their implementation and have more market power or political influence than those who would win.

The distinction between differences in technology and differences in efficiency is often unclear. Suppose firms in country A are using the same machinery and the same number of workers per machine as in country B, but output per worker is higher in A than B. This may appear to be an obvious case of inefficiency, since the technology embodied in the machines used by workers in the two countries is the same. But maybe it is just that people in country B lack the knowledge of how best to use the machines, in which case it may actually be a case of differences in technology. As an example, General Motors has had little success in their attempts to emulate the manufacturing

methods that Toyota has deployed successfully for many years even in their US operations.

Moreover, identical technologies will have different effects in different countries, because of differences in language, raw materials, consumer preferences, workers expectations and the like. Euro Disney, for example, was plagued initially with labour disputes when it first opened its park in the outskirts of Paris in 1987. It took the American managers several years to realize that the problem was not recalcitrant workers but rather that French workers consider it an intolerable indignity to be forced to wear items such as mouse ears when serving the public. A minor adjustment in amusement park technology was needed to make it as productive in France as it had been in the United States.

Deeper Determinants of Growth

Even if we knew how much of the cross-country variation in growth rates or income levels to attribute to different kinds of capital or to technology or efficiency, we would still be faced with the deeper question of why these differences in capital and productivity arise. A large number of candidate explanations have been offered in the literature. These candidates can be classified into four broad categories: geography, institutions, policy and culture.

Geographical differences are perhaps the most obvious. As Sachs (2003) has emphasized, countries that are landlocked, that suffer from a hazardous disease environment and that have difficult obstacles in the way of internal transport, will almost certainly produce at a lower level than countries without these problems, even if they use the same technology and the same array of capital. In addition, the lower productivity of these countries will serve to reduce the rate of return to accumulating capital and to generating new technologies.

Institutions matter because of the way they affect private contracts and also because of the way they affect the extent to which the returns to different kinds of investments can be appropriated by the government. The origin of a country's legal system has been shown by La Porta et al. (1998) to have an important effect on private contracts. In particular, these authors show that countries with British legal origins tend of offer greater protection of investor and creditor rights, which in turn is likely to affect both capital accumulation and investment in technology by making outside finance more easily available.

Because long-term productivity growth requires technical progress, it depends on political, institutional and regulatory factors that affect the way the conflict between the winners and losers of technical progress will be resolved, and hence affect the incentives to create and adopt new technologies. For example, the way intellectual property is protected will affect the incentive to innovate, because on the one hand no one will want to spend resources creating new technologies that his or her rivals can easily copy, while on the other hand a firm that is protected from competition by patent laws that make it difficult for rivals to innovate in the same product lines will be under less pressure to innovate. Likewise, a populist political regime may erect barriers to labour-saving innovation, resulting in slower technical progress.

Economic policies matter not only because of the way they affect the return to investing in capital and technology but also because of the inefficiencies that can be created by taxes and subsidies. But how these policies affect economic growth can vary from one country to another. In particular, Aghion and Howitt (2006) have argued that growth-promoting policies in technologically advanced countries are not necessarily growth-promoting in poorer countries, because innovation and implementation are affected differently by the same variables. For example, tighter competition policy in a relatively backward country might retard technology development by local firms that will be discouraged by the threat of foreign entry, whereas in more advanced countries firms will be spurred on to make even greater R&D investments when threatened by competition.

As this example suggests, international trade is one of the policy domains most likely to matter for growth and income differences, because of the

huge productivity advantage that is squandered by policies that run counter to comparative advantage, because protected firms tend to become technologically backward firms, and because for many countries international trade is the only way for firms to gain a market large enough to cover the expense of developing leading-edge technologies. So it is probably no accident that export promotion has been a prominent feature of all the East Asian countries that began escaping from the lower end of the world income distribution towards the end of the 20th century, whereas import substitution was a prominent feature of several Latin American countries that fell from the upper end of the distribution early in the 20th century.

Culture is a difficult factor to measure. In principle, however, it is capable of explaining a great deal of cross-country variation in growth, because a society in which people are socialized to trust each other, to work hard, to value technical expertise and to respect law and order is certainly going to be thriftier and more productive than a society in which these traits do not apply. Recent work has begun to quantify the role of culture using measures of social capital, social capability, ethnolinguistic fractionalization, religious belief, the spread of Anglo-Saxon culture and many other variables.

See Also

- ▶ Economic Growth, Empirical Regularities in
- ▶ Economic Growth in the Very Long Run
- ▶ Endogenous Growth Theory
- ▶ Growth Accounting
- ▶ Growth and Institutions
- ▶ Growth Take-Offs
- ▶ Level Accounting
- ▶ Population Ageing

Bibliography

Acemoglu, D., and F. Zilibotti. 2001. Productivity differences. *Quarterly Journal of Economics* 116: 563–606.

Aghion, P., and P. Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60: 323–351.

Aghion, P., and P. Howitt. 2006. Appropriate growth policy: An integrating framework. *Journal of the European Economic Association* 4: 269–314.

Aghion, P., P. Howitt, and D. Mayer-Foulkes. 2005. The effect of financial development on convergence: Theory and evidence. *Quarterly Journal of Economics* 120: 173–222.

Basu, S., and D.N. Weil. 1998. Appropriate technology and growth. *Quarterly Journal of Economics* 113: 1025–1054.

Bourguignon, F., and C. Morrison. 2002. Inequality among world citizens: 1820–1992. *American Economic Review* 92: 727–744.

Caselli, F. 2005. Accounting for cross-country income differences. In *Handbook of economic growth*, ed. P. Aghion and S.N. Durlauf, vol. 1. Amsterdam: North-Holland.

Evans, P. 1996. Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* 20: 1027–1049.

Evenson, R.E., and L.E. Westphal. 1995. Technological change and technology strategy. In *Handbook of development economics*, ed. T.N. Srinivasan and J. Behrman, vol. 3A. Amsterdam: Elsevier.

Fogel, R. 1997. New findings on secular trends in nutrition and mortality: Some implications for population theory. In *Handbook of population and family economics*, ed. M.R. Rosenzweig and O. Stark, vol. 1A. Amsterdam: North-Holland.

Frankel, M. 1962. The production function in allocation and growth: A synthesis. *American Economic Review* 52: 995–1022.

Galor, O., and D.N. Weil. 2000. Population, technology, and growth: From Malthusian stagnation to the demographic transition and beyond. *American Economic Review* 90: 806–828.

Gerschenkron, A. 1952. Economic backwardness in historical perspective. In *The Progress of Underdeveloped Areas*, ed. B.F. Hoselitz. Chicago: University of Chicago Press.

Gollin, D. 2002. Getting income shares right. *Journal of Political Economy* 110: 458–474.

Greenwood, J., and Yorukoglu, M. 1997. 1974. *Carnegie-Rochester Conference Series on Public Policy* 46: 49–95.

Hall, R., and C. Jones. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114: 83–116.

Howitt, P. 1998. Measurement, obsolescence, and general purpose technologies. In *General purpose technologies and economic growth*, ed. E. Helpman. Cambridge, MA: MIT Press.

Howitt, P. 2000. Endogenous growth and cross-country income differences. *American Economic Review* 90: 829–846.

Howitt, P., and P. Aghion. 1998. Capital accumulation and innovation as complementary factors in long-run growth. *Journal of Economic Growth* 3: 111–130.

Hsieh, C.-T., and P. Klenow. 2007. Relative prices and relative prosperity. *American Economic Review* 97.
Jones, C.I. 1997. On the evolution of the world income distribution. *Journal of Economic Perspectives* 11(3): 19–36.
Klenow, P., and A. Rodriguez-Clare. 1997. The neoclassical revival in growth economics: Has it gone too far? In *NBER macro annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.
La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R.W. Vishny. 1998. Law and finance. *Journal of Political Economy* 106: 1113–1155.
Lipsey, R.G., K.I. Carlaw, and C.T. Bekar. 2005. *Economic transformations: General purpose technologies and long term economic growth*. New York: Oxford University Press.
Lucas, R.E.Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.
Maddison, A. 2001. *The world economy: A millennial perspective,* Development centre studies. Paris: OECD.
Malthus, T.R. 1798. *An essay on the principle of population, as it affects the future improvement of society with remarks on the speculations of Mr. Godwin, M. Condorcet, and other writers.* London: printed for J. Johnson in St Paul's Churchyard.
Mayer-Foulkes, D. 2002. *Global divergence. Documento de Trabajo del CIDE, SDTE 250, División de Economía*. Mexico: CIDE.
Pritchett, L. 1997. Divergence, big-time. *Journal of Economic Perspectives* 11(3): 3–17.
Rodriguez, F., and J.D. Sachs. 1999. Why do resource-abundant economies grow more slowly? *Journal of Economic Growth* 4: 277–303.
Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037.
Romer, P.M. 1990. Endogenous technological change. *Journal of Political Economy* 98: S71–102.
Rosenberg, N. 1981. How exogenous is science? In *Inside the black box: Technology and economics*, ed. N. Rosenberg. New York: Cambridge University Press.
Sachs, J.D. 2003. *Institutions don't rule: Direct effects of geography on per capita income* (Working Paper No. 9490). Cambridge, MA: NBER.
Sala-i-Martin, X. 2006. The world distribution of income: Falling poverty and . . . convergence, period. *Quarterly Journal of Economics* 121: 351–397.
Schumpeter, J.A. 1942. *Capitalism, socialism and democracy.* New York: Harper.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39: 312–320.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Weil, D.N. 2005. *Economic growth*. Boston: Addison-Wesley.
Weil, D.N. 2007. Accounting for the effect of health on economic growth. *Quarterly Journal of Economics* 122: 1265–1306.

Economic Growth in the Very Long Run

Oded Galor

E

Abstract

The evolution of economies during the major portion of human history was marked by Malthusian stagnation. The transition from an epoch of stagnation to a state of sustained economic growth has shaped the contemporary world economy and has led to the great divergence in income per capita across the globe in the past two centuries. This article examines the process of development over the course of human history in light of recent advances in unified growth theory.

Keywords

Capital accumulation; China; Colonialism; Convergence clubs; Demographic transition; Economic growth in the very long run; Education; Endogenous growth; Exogenous growth; Fertility; Globalization; Great divergence; Growth take-offs; Human capital; Income growth; Industrial Revolution; Industrialization; Inequality; Land; Land productivity; Literacy; Malthusian stagnation; Marriage; Microfoundations; Population growth; Post-Malthusian regime; Sustained growth regime; Technological progress; Unified growth theory

JEL classifications

N1; O4; O11; O14; O33; O40; J11; J13

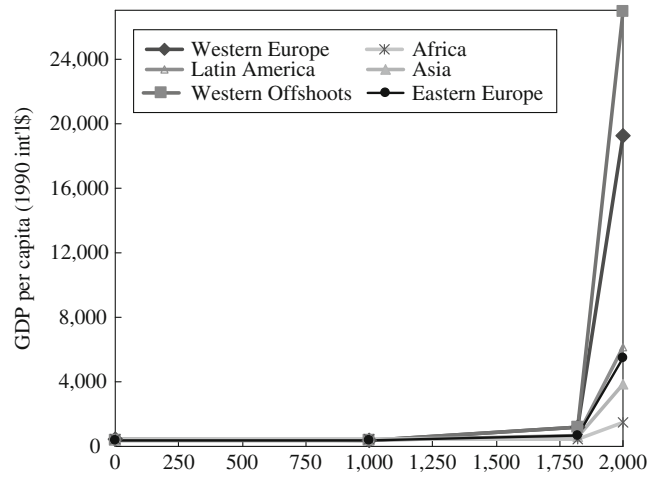
The evolution of economies during the major portion of human history was marked by Malthusian stagnation. Technological progress and population growth were minuscule by modern standards, and

the average growth rates of income per capita in various regions of the world were even slower due to the offsetting effect of population growth on the expansion of resources per capita.

In the past two centuries the pace of technological progress increased significantly in association with the process of industrialization. Various regions of the world departed from the Malthusian trap and experienced a considerable rise in the growth rates of income per capita and population. Unlike episodes of technological progress in the pre-Industrial Revolution era that failed to generate sustained economic growth, the increasing role of human capital in the production process in the second phase of industrialization ultimately prompted a demographic transition, liberating the gains in productivity from the counterbalancing effects of population growth. The decline in the growth rate of population and the enhancement of human capital formation and technological progress paved the way for the emergence of the modern state of sustained economic growth. Variations in the timing of the transitions from a Malthusian epoch to a state of sustained economic growth across countries lead to a considerable rise in the ratio of GDP per capita between the richest and the poorest regions of the world from 3:1 in 1820 to 18:1 in 2000 (see Fig. 1).

The transition from stagnation to growth and the associated phenomenon of the great divergence have been the subject of intensive research in the growth literature in recent years (Galor and Weil 1999, 2000; Galor and Moav 2002; Lucas 2002; Hansen and Prescott 2002; Jones 2001; Hazan and Berdugo 2002; Doepke 2004; Lagerlof 2003, 2006; Galor and Mountford 2003, 2006). The inconsistency of exogenous and endogenous growth models with some of the most fundamental features of the process of development has led to a search for a unified theory that would unveil the underlying microfoundations of the growth process in its entirety, and would capture in a single framework the epoch of Malthusian stagnation that characterized most of human history, the contemporary era of modern economic growth, and the driving forces that triggered the recent transition between these regimes.

The advance of unified growth theory was fuelled by the conviction that the understanding of the contemporary growth process would be fragile and incomplete unless growth theory were based on proper microfoundations that reflect the various qualitative aspects of the growth process and their central driving forces. Moreover, it has become apparent that a comprehensive understanding of the hurdles faced by less developed economies in reaching a state of sustained economic growth would remain obscure unless the factors that prompted the transition of the currently developed economies into a state of sustained economic growth could be identified and modified to account for the differences in the growth structure of less developed economies in an interdependent world.



Economic Growth in the Very Long Run, Fig. 1 The evolution of regional income per capita, 1–2000 (Source: Maddison (2001))

Unified growth theory explores the fundamental factors that generated the remarkable escape from the Malthusian epoch and their significance in understanding the contemporary growth process of developed and less developed economies. Moreover, it sheds light on the perplexing phenomenon of the great divergence in income per capita across regions of the world in the past two centuries. It suggests that the transition from stagnation to growth is an inevitable outcome of the process of development. The inherent Malthusian interaction between the level of technology and the size and the composition of the population accelerated the pace of technological progress and ultimately raised the importance of human capital in the production process. The rise in the demand for human capital in the second phase of industrialization and its impact on the formation of human capital as well as on the onset of the demographic transition brought about significant technological advances along with a reduction in fertility rates and population growth, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita, and paving the way for the emergence of sustained economic growth.

Differences in the timing of the take-off from stagnation to growth across countries (for example, England's earlier industrialization in comparison with China) contributed significantly to the great divergence and to the emergence of convergence clubs. These variations reflect initial differences in geographical factors and historical accidents and their manifestation in diversity in institutional, demographic, and cultural factors, trade patterns, colonial status, and public policy. In particular, once a technologically driven demand for human capital emerged in the second phase of industrialization, the prevalence of human capital-promoting institutions determined the extensiveness of human capital formation, the timing of the demographic transition, and the pace of the transition from stagnation to growth. Thus, unified growth theory provides the natural framework of analysis in which variations in the economic performance across countries and regions could be examined based on the effect of variations in educational, institutional, geographical, and cultural factors on the pace of the transition from stagnation to growth.

The Process of Development

The process of economic development has been characterized by of three fundamental regimes: the Malthusian epoch, the post-Malthusian regime, and the sustained growth regime.

The Malthusian Epoch

During the Malthusian epoch that characterized most of human history, humans were subjected to a persistent struggle for existence. Resources generated by technological progress and land expansion were channeled primarily towards an increase in the size of the population, with a minor long-run effect on income per capita. Improvements in the technological environment or in the availability of land generated temporary gains in income per capita, leading eventually to a larger but not richer population. Technologically superior countries ultimately had denser populations but their standard of living did not reflect the degree of their technological advancement.

During the Malthusian epoch the average growth rate of output per capita was negligible and the standard of living did not differ greatly across countries. The average level of income per capita in the world during the first millennium fluctuated around $450 per year (in 1990 international dollars) and the average growth rate of output per capita was nearly zero (Maddison 2001). This state of Malthusian stagnation persisted until the end of the 18th century. In the years 1000–1820, the average level of income per capita in the world economy was below $670 per year, and the average growth rate of the world income per capita was minuscule, creeping at a rate of about 0.05 per cent per year. Nevertheless, income per capita fluctuated significantly within regions, deviating from their sluggish long-run trend over decades and sometimes centuries.

Population growth over this era followed the Malthusian pattern as well. The gradual increase in income per capita during the Malthusian epoch was

associated with a monotonic increase in the average rate of growth of world population. The slow pace of resource expansion in the first millennium was reflected in a modest increase in the population of the world from 231 million people in 1 CE to 268 million in 1000 CE: a minuscule average growth rate of 0.02 per cent per year. The more rapid (but still very slow) expansion of resources in the period 1000–1500 permitted the world population to increase by 63 per cent, from 268 million in 1000 to 438 million in 1500; a slow 0.1 per cent average growth rate per year. Resource expansion over the period 1500–1820 had a more significant impact on the world population, which grew 138 per cent from 438 million in 1500 to 1041 million in 1820: an average pace of 0.27 per cent per year.

Variations in population density across countries during the Malthusian epoch reflected primarily cross-country differences in technology and land productivity. Due to the positive adjustment of the population to an increase in income per capita, differences in technology or in land productivity across countries resulted in variations in population density rather than in the standard of living. For instance, China's technological advancement in the period 1500–1820 permitted its share of world population to increase from 23.5 per cent to 36.6 per cent, while its income per capita in the beginning and the end of this time interval remained approximately $600 per year.

The Post-Malthusian Regime

During the post-Malthusian regime, the pace of technological progress markedly increased in association with the process of industrialization, triggering a take-off from the Malthusian trap. The growth rate of income per capita increased significantly but the positive Malthusian effect of income per capita on population growth was still maintained, generating a sizeable increase in population growth that offset some of the potential gains in income per capita.

The take-off of developed regions from the Malthusian regime occurred at the beginning of the 19th century and was associated with the Industrial Revolution, whereas the take-off of less developed regions occurred towards the beginning of the 20th century and was delayed in some countries well into the 20th century. During the post-Malthusian regime the average growth rate of output per capita increased significantly and the standard of living began to differ considerably across countries. The average growth rate of output per capita in the world soared from 0.05 per cent per year during the period 1500–1820 to 0.53 per cent per year in the years 1820–70, and 1.3 per cent per year during the period 1870–1913. The timing of the take-off and its magnitude differed across regions. The take-off from the Malthusian epoch and the transition to the post-Malthusian regime occurred in western Europe, the Western offshoots (that is, the United States, Canada, Australia and New Zealand), and eastern Europe at the beginning of the 19th century, whereas in Latin America, Asia and Africa it occurred towards the beginning of the 20th century.

The rapid increase in income per capita in the post-Malthusian regime was channeled partly towards an increase in the size of the population. During this period, the Malthusian mechanism linking higher income to higher population growth continued to function. However, the effect of higher population on the dilution of resources per capita was counteracted by accelerated technological progress and capital accumulation, allowing income per capita to rise despite the offsetting effects of population growth.

The western European take-off along with that of the Western offshoots brought about a sharp increase in population growth in these regions and consequently a modest rise in population growth in the world as a whole. The subsequent take-off of less developed regions, and the associated increase in their rates of population growth, brought about a significant rise in population growth in the world. The rate of population growth in the world increased from an average rate of 0.27 per cent per year in the period 1500–1820 to 0.4 per cent per year in the years 1820–70, and to 0.8 per cent per year in the time interval 1870–1913. Despite the decline in population growth in western Europe and the Western offshoots towards the end of the 19th century and the beginning of the 20th century, the delayed take-off of less developed regions, and the

significant increase in their income per capita prior to their demographic transitions, generated a further increase in the rate of population growth in the world to 0.93 per cent per year in the period 1913–50, and 1.92 per cent per year in the period 1950–73. Ultimately, the onset of the demographic transition in less developed economies during the second half of the 20th century reduced population growth rates to 1.66 per cent per year in the 1973–98 period (Maddison 2001).

It appears that the significant rise in income per capita in the post-Malthusian regime increased the desired number of surviving offspring and thus, despite the decline in mortality rates, fertility increased significantly so as to enable households to reach this higher desired level of surviving offspring. Fertility was controlled during this period, despite the absence of modern contraceptive methods, partly via adjustment in marriage rates. Increased fertility was achieved by earlier female age of marriage, and a decline in fertility by a delay in the marriage age.

The take-off in the developed regions was accompanied by a rapid process of industrialization. Per-capita level of industrialization increased significantly in the United Kingdom, rising 50 per cent over the 1750–1800 period, quadrupling in the years 1800–60, and nearly doubling in the time period 1860–1913. Similarly, per capita level of industrialization accelerated in the United States, doubling in the 1750–1800 as well as 1800–60 periods, and increasing sixfold in the years 1860–1913. A similar pattern was experienced in Germany, France, Sweden, Switzerland, Belgium and Canada. The take-off of less developed economies in the 20th century was associated with increased industrialization as well. However, during the 19th century these economies experienced a decline in per capita industrialization, reflecting the adverse effect of the sizeable increase in population on the level of industrial production per capita as well as the forces of globalization and colonialism, which induced less developed economies to specialize in the production of raw materials (Galor and Mountford 2003, 2006).

The acceleration in technological progress during the post-Malthusian regime and the associated increase in income per capita stimulated the accumulation of human capital in the form of literacy rates, schooling, and health. The increase in the investment in human capital was induced by the rise in income per capita, as well as by qualitative changes in the economic environment that increased the demand for human capital and induced households to invest in the education of their offspring.

In the first phase of the Industrial Revolution, human capital had a limited role in the production process. Education was motivated by a variety of reasons, such as religion, enlightenment, social control, moral conformity, socio-political stability, social and national cohesion, and military efficiency. The extensiveness of public education was therefore not necessarily correlated with industrial development, and it differed across countries due to political, cultural, social, historical and institutional factors. In the second phase of the Industrial Revolution, however, the demand for education increased, reflecting the increasing skill requirements in the process of industrialization. The economic interests of capitalists were a significant driving force behind the implementation of educational reforms (Galor and Moav 2006). The process of industrialization has been characterized by a gradual increase in the relative importance of human capital in less developed economies as well and educational attainment increased significantly across all less developed regions in the post-Malthusian regime.

The Sustained Growth Regime

The acceleration in the rate of technological progress in the second phase of industrialization, and its interaction with human capital formation, triggered a demographic transition, paving the way to a transition to an era of sustained economic growth. In the post demographic-transition period, the rise in aggregate income due to technological progress and factors accumulation was no longer counterbalanced by population growth, permitting sustained growth in income per capita in regions that experienced sustained technological progress and accumulation of physical and human capital.

The transition of the developed regions of western Europe and the Western offshoots to the state of sustained economic growth occurred

towards the end of the 19th century, and their income per capita in the 20th century has advanced at a stable rate of about two per cent per year. The transition of some less developed countries in Asia and Latin America occurred towards the end of the 20th century. Africa, in contrast, is still struggling to make this transition.

The transition to a state of sustained economic growth was characterized by a gradual increase in the importance of the accumulation of human capital relative to physical capital as well as with a sharp decline in fertility rates. In the first phase of the Industrial Revolution (1760–1830), capital accumulation as a fraction of GDP significantly increased whereas literacy rates remained largely unchanged. Skills and literacy requirements were minimal, the state devoted virtually no resources to raise the level of literacy of the masses, and workers developed skills primarily through on-the-job training (Green 1990; Mokyr 1993). Consequently, literacy rates did not increase during the period 1750–1830 (Sanderson 1995).

In the second phase of the Industrial Revolution, however, the pace of capital accumulation subsided, skills became necessary for production and the education of the labour force markedly increased. The investment ratio in the UK, which increased from six per cent in 1760 to 11.7 per cent in 1831, remained at around 11 per cent on average in the years 1856–1913 (Crafts 1985). In contrast, the average years of schooling of males in the labour force that did not change significantly until the 1830s tripled by the beginning of the 20th century. The drastic rise in the level of income per capita in England as of 1865 was associated with an increase in school enrolment of ten-year-old children from 40 per cent in 1870 to 100 per cent in 1900. Moreover, total fertility rate in England sharply declined over this period from about five in 1875, to nearly two in 1925.

The demographic transition swept the world in the course of the 20th century. The unprecedented increase in population growth during the post-Malthusian regime was reversed and the demographic transition brought about a significant reduction in fertility rates and population growth in various regions of the world, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita. The demographic transition enhanced the growth process via three channels: (*a*) reductions in the dilution of the stocks of capital and natural resources, (*b*) enhancements in human capital formation, and (*c*) changes in the age distribution of the population, temporarily increasing the size of the labour force relative to the population as a whole.

The timing of the demographic transition differed significantly across regions. The reduction in population growth occurred in Western Europe, the Western offshoots, and eastern Europe towards the end of the 19th century and in the beginning of the 20th century, whereas Latin America and Asia experienced a decline in the rate of population growth only in the last decades of the 20th century. Africa's population growth, in contrast, has been rising steadily.

The process of industrialization was characterized by a gradual increase in the relative importance of human capital in the production process. The acceleration in the rate of technological progress gradually increased the demand for human capital, inducing individuals to invest in education, and stimulating further technological advancement. Moreover, in developed as well as less developed regions, the onset of the process of human capital accumulation preceded the onset of the demographic transition, suggesting that the rise in the demand for human capital in the process of industrialization and the subsequent accumulation of human capital played a significant role in the demographic transition and the shift to a state of sustained economic growth.

Notably, the reversal of the Malthusian relation between income and population growth during the demographic transition corresponded to an increase in the level of resources invested in each child. For example, literacy rate among men in England was stable at around 65 per cent in the first phase of the Industrial Revolution and increased significantly during the second phase, reaching nearly 100 per cent at the end of the 19th century. In addition, the proportion of children aged 5 to 14 in primary schools increased from 11 per cent in 1855 to 74 per cent in 1900.

A similar pattern is observed in other European societies (Flora, Kraus and Pfenning 1983).

The process of industrialization was characterized by a gradual increase in the relative importance of human capital in less developed economies as well. Educational attainment increased significantly across all less developed regions. Moreover, in line with the pattern that emerged among developed economies in the 19th century, the increase in educational attainment preceded or occurred simultaneously with the decline in total fertility rates.

The Great Divergence

The differential timing of the take-off from stagnation to growth across countries and the corresponding variations in the timing of the demographic transition led to a great divergence in income per capita as well as population growth. Inequality in the world economy was negligible till the 19th century. The ratio of GDP per capita between the richest region and the poorest region in the world was only 1.1:1 in 1000, 2:1 in 1500 and 3:1 in 1820. In the past two centuries, however, the ratio of GDP per capita between the richest group (Western offshoots) and the poorest region (Africa) has widened considerably from a modest 3:1 ratio in 1820, to 5:1 ratio in 1870, 9:1 ratio in 1913, 15:1 ratio in 1950, and 18:1 ratio in 2001.

An equally momentous transformation occurred in the distribution of world population across regions. The earlier take-off of western European countries increased the amount of resources that could be devoted for the increase in family size, permitting a 16 per cent increase in the share of their population in the world from 12.8 per cent in 1820 to 14.8 per cent in 1870. However, the early onset in the western European demographic transition and the long delay in the demographic transition of less developed regions, well into the second half of the 20th century, led to a decline in the share of western European population in the world, from 14.8 per cent in 1870 to 6.6 per cent in 1998. In contrast, the prolongation of the post-Malthusian period among less developed regions, in association with the delay in their demographic transition well into the second half of 20th century, channelled their increased resources towards a significant increase in their population. Africa's share of world population increased from seven per cent in 1913 to 12.9 per cent in 1998, Asia's share of world population increased from 51.7 per cent in 1913 to 57.4 per cent in 1998, and Latin American countries increased their share in world population from two per cent in 1820 to 8.6 per cent in 1998.

E

Unified Growth Theory

Galor and Weil (2000) advanced a unified growth theory that captures the three regimes that have characterized the process of development as well as the fundamental driving forces that generated the transition from an epoch of Malthusian stagnation to a state of sustained economic growth. The theory replicates the observed time paths of population, income per capita, and human capital, generating: (*a*) the Malthusian oscillations in population and output per capita during the Malthusian epoch, (*b*) an endogenous take-off from Malthusian stagnation that is associated with an acceleration in technological progress and is accompanied initially by a rapid increase in population growth, and (*c*) a rise in the demand for human capital, followed by a demographic transition and sustained economic growth. These qualitative patterns are confirmed in the calibration of the theory by Lagerlof (2006).

The theory proposes that in early stages of development economies were in the proximity of a stable Malthusian equilibrium. Technology advanced rather slowly, and generated proportional increases in output and population. The inherent positive interaction between population and technology in this epoch, however, gradually increased the pace of technological progress, and due to the delayed adjustment of population, output per capita advanced at a minuscule rate. The slow pace of technological progress in the Malthusian epoch provided a limited scope for human capital in the production process and parents, therefore, had no incentive to reallocate resources towards human capital formation of their offspring.

The Malthusian interaction between technology and population accelerated the pace of technological progress and permitted a take-off to the post-Malthusian regime. The expansion of resources was partially counterbalanced by the enlargement of population, and the economy was characterized by rapid growth rates of income per capita and population. The acceleration in technological progress eventually increased the demand for human capital, generating two opposing effects on population growth. On the one hand, it eased households' budget constraints, allowing the allocation of more resources for raising children. On the other hand, it induced a reallocation of resources towards child quality. In the post-Malthusian regime, due to the modest demand for human capital, the first effect dominated, and the rise in real income permitted households to increase the number as well the quality of their children.

As investment in human capital took place, the Malthusian steady-state equilibrium vanished and the economy started to be attracted by the gravitational forces of the modern growth regime. The interaction between investment in human capital and technological progress generated a virtuous circle: human capital generated faster technological progress, which in turn further raised the demand for human capital, inducing further investment in child quality, and eventually triggering the onset of the demographic transition and the emergence of a state of sustained economic growth.

The theory suggests that the transition from stagnation to growth is an inevitable outcome of the process of development. The inherent Malthusian interaction between the level of technology and the size of the population accelerated the pace of technological progress, and ultimately raised the importance of human capital in the production process. The rise in the demand for human capital in the second phase of the Industrial Revolution and its impact on the formation of human capital as well as on the onset of the demographic transition brought about significant technological advancements along with a reduction in fertility rates and population growth, enabling economies to convert a larger share of the fruits of factor accumulation and technological progress into growth of income per capita, and paving the way for the emergence of sustained economic growth. Quantitative analysis of unified growth theories (Doepke 2004); Lagerlof 2006) indeed suggest that the rise in the demand for human capital was a significant force behind the demographic transition and the emergence of a state of sustained economic growth.

Variations in the timing of the transition from stagnation to growth and thus in economic performance across countries reflect initial differences in geographical factors and historical accidents and their manifestation in diversity in institutional, demographic, and cultural factors, trade patterns, colonial status, and public policy. In particular, once a technologically driven demand for human capital emerged in the second phase of industrialization, the prevalence of human capital-promoting institutions determined the extensiveness of human capital formation, the timing of the demographic transition, and the pace of the transition from stagnation to growth.

The theory proposes that the growth process is characterized by stages of development and it evolves nonlinearly. Technological leaders experienced a monotonic increase in the growth rates of their income per capita. Their growth was rather slow in early stages of development, increased rapidly during the take-off from the Malthusian epoch, and continued to rise, often stabilizing at higher levels. In contrast, technological followers that made the transition to sustained economic growth experienced a non-monotonic increase in the growth rates of their income per capita. Their growth rate was rather slow in early stages of development, but increased rapidly in the early stages of the take-off from the Malthusian epoch, boosted by the adoption of technologies from the existing technological frontier. However, once these economies reached the technological frontier, their growth rates dropped to the level of the technological leaders. Hence, consistently with contemporary evidence about the existence of multiple growth regimes (Durlauf and Quah 1999), the differential timing of the take-off from stagnation to growth across economies generated convergence clubs characterized

by a group of poor countries in the vicinity of the Malthusian equilibrium, a group of rich countries in the vicinity of the sustained growth equilibrium, and a third group in the transition from one club to another.

See Also

- ▶ Growth Take-Offs
- ▶ Human Capital, Fertility and Growth

Bibliography

Crafts, N.F.R. 1985. *British economic growth during the industrial revolution*. Oxford: Oxford University Press.

Doepke, M. 2004. Accounting for fertility decline during the transition to growth. *Journal of Economic Growth* 9: 347–383.

Durlauf, S.N., and D. Quah. 1999. The new empirics of economic growth. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Flora, P., F. Kraus, and W. Pfenning. 1983. *State, Economy and Society in Western Europe 1815–1975*. Chicago: St. James Press.

Galor, O. 2005. From stagnation to growth: unified growth theory. In *Handbook of economic growth*, ed. P. Aghion and S.N. Duraluf. Amsterdam: North-Holland.

Galor, O., and O. Moav. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117: 1133–1192.

Galor, O., and O. Moav. 2006. Das human kapital: A theory of the demise of the class structure. *Review of Economic Studies* 73: 85–117.

Galor, O., and A. Mountford . 2003. *Trading population for productivity.* Working paper, Brown University.

Galor, O., and A. Mountford. 2006. Trade and the great divergence: The family connection. *American Economic Review* 96: 299–303.

Galor, O., and D.N. Weil. 1999. From Malthusian stagnation to modern growth. *American Economic Review* 89: 150–154.

Galor, O., and D.N. Weil. 2000. Population, technology and growth: from the Malthusian regime to the demographic transition and beyond. *American Economic Review* 110: 806–828.

Green, A. 1990. *Education and state formation*. New York: St. Martin's Press.

Hansen, G., and E. Prescott. 2002. Malthus to Solow. *American Economic Review* 92: 1205–1217.

Hazan, M., and B. Berdugo. 2002. Child labor, fertility and economic growth. *Economic Journal* 112: 810–828.

Jones, C.I. 2001. Was an industrial revolution inevitable? Economic growth over the very long run. *Advances in Macroeconomics* 1: 1–43.

Lagerlof, N. 2003. From Malthus to modern growth: the three regimes revisited. *International Economic Review* 44: 755–777.

Lagerlof, N. 2006. The Galor–Weil model revisited: a quantitative exploration. *Review of Economic Dynamics* 9: 116–142.

Lucas, R.E. 2002. *The industrial revolution: Past and future*. Cambridge, MA: Harvard University Press.

Maddison, A. 2001. *The world economy: A millennia perspective*. Paris: OECD.

Mokyr, J. 1993. The new economic history and the industrial revolution. In *The British industrial revolution: An economic perspective*, ed. J. Mokyr. Boulder, CO: Westview Press.

Sanderson, M. 1995. *Education, Economic Change and Society in England 1780–1870*. Cambridge: Cambridge University Press.

Economic Growth Non-linearities

Chih Ming Tan

Abstract

Nonlinearities in growth have important implications for cross-country income inequality. In particular, they imply that countries may spend long periods of time in a low-growth poverty trap. However, finding evidence of such nonlinearities in the data and accounting for their emergence pose unique challenges to researchers.

Keywords

Balanced growth; Conditional convergence; Convergence; Convergence clubs; Demographic transition; Diffusion of technology; Economic development; Economic growth nonlinearities; Increasing returns; Industrialization; Market size; Multiple-growth regimes; Neoclassical growth theory; Neoclassical production function; Nonconvexity; Poverty traps; Spillover effects; Stages of theory of growth; Strategic complementarities; Structural change; Take-off; Technological progress

JEL Classifications

O4

Nonlinear Growth Models

Nonlinear growth models are characterized by a country's subsequent performance being critically dependent upon its initial conditions. In particular, these models tend to imply that countries which have unfavourable initial conditions may either experience substantial periods of time in low-growth/low-income poverty traps or be altogether caught in one. In some cases, it has been explicitly suggested that active (exogenous) policy interventions may be necessary in order to kick-start a country into a more favourable equilibrium. Nonlinear growth models can be broadly classified into two classes: structural change (or 'stages of development') models, and models that emphasize endogenous technological development and cross-country interactions in terms of technological diffusion.

Structural change models focus on the (internal) transformations of an economy as it transits through critical phases or 'stages' (see Lewis 1956; Rostow 1960) leading to industrialization. The aim of this work is to clarify the conditions for such transitions to occur. Early work in the economic development literature (see Rosenstein-Rodan 1943; Nurkse 1953; Scitovsky 1954; Fleming 1955; formalized by Murphy et al. 1989) emphasized the importance of increasing returns and the size of the market in industrialization. The key idea behind this view is that countries could be locked in a no-industrialization trap because of the small size of the market for each sector of the economy. No single sector can achieve growth on its own. However, the growth of one sector results in the enlargement of markets for other sectors. The enlargement of markets then encourages investment and growth in the corresponding sectors. These spillover effects and strategic complementarities imply that a 'big push' – that is, coordinated investments (or 'balanced growth') across sectors – may be sufficient to push the economy out of the trap and into a 'take-off' towards industrialization. Other models are explicitly informed by the analysis of historical data (see Maddison 2004), and emphasize the importance of explaining simultaneously both historical patterns of other state variables associated with growth and growth itself. An important recent work that models the demographic transition in growth take-offs is Galor and Weil (2000). Because these models require that certain conditions be met before countries are able to achieve take-off, those who do not meet these requirements could find themselves trapped in a phase of economic stagnation for extended periods of time.

The second class of models focuses on the role of technological progress in growth. In particular, the emphasis of these models is on the diffusion of technology from countries which are technological leaders to less developed countries. Lucas (2000) is a seminal work in this area (see also Basu and Weil 1998; Parente and Prescott 1994; Howitt and Mayer-Foulkes 2005). Particular attention has been paid to exploring the channels through which less advanced countries imitate or adopt technologies in leader countries. If there are no barriers to technological diffusion across countries, then these models typically predict that rich and poor countries would gradually converge in per capita income. However, if such barriers exist, then countries may differ in their ability to adopt technologies leading to the creation of 'clusters' of countries defined by a set of common barriers to technological adoption. Countries within each of these clusters or 'convergence clubs' converge to common levels of mean per capita income. Nevertheless, the per capita incomes across convergence clubs need never converge and the polarization of per capita incomes across countries may be permanent.

Growth Empirics

In both classes of models, therefore, the primary concern is that countries may become separated – perhaps permanently – into multiple growth regimes corresponding to different levels of long-run per capita income. The fact that nonlinear growth models imply that global inequality may be persistent has sparked major advances in the area of cross-country growth empirics. Driven by such concerns, the central preoccupation of growth empirics has been to evaluate the

conditions under which poor countries catch up with rich ones or fail to do so. Initial work along these lines focused on the concept of 'conditional convergence'. Conditional convergence is said to occur if permanent per capita income differences between countries can be accounted for solely by structural differences (and not initial conditions). Researchers initially argued that because conditional convergence was predicted by the canonical neoclassical growth model (see Ramsey 1928; Solow 1956; Swan 1956; Cass 1965; Koopmans 1965) whereas nonlinear growth models potentially predict dependence on initial conditions, tests for conditional convergence could be used to discriminate between these classes of theories.

Following Mankiw et al. (1992) and Barro and Sala-i-Martin (1992), the canonical way such tests were conducted was to first construct a linearized version of the neoclassical growth model about the (unique) steady state with average growth rates across a time period as the dependent variable, and measures of physical and human capital, population growth rates, and initial per capita income as covariates. Researchers then applied the linearized neoclassical model to cross-country data with the aim of testing to see whether the data supported a negative coefficient on initial per capita income. A finding of a negative coefficient on initial per capita income was taken to imply that, conditional on countries having similar structural characteristics (as defined by the set of covariates), poorer countries would close the income gap with the rich – that is, conditional convergence.

An important outcome of the, oftentimes heated (see Sala-i-Martin 1996), convergence debates of the 1990s was precisely to weaken the idea that such tests of convergence could be interpreted as model selection tests. In a highly influential work, Bernard and Durlauf (1996) strongly disputed the interpretation of such 'conditional convergence' tests by pointing out that these tests were not able to discriminate against a class of nonlinear growth theories that have dramatically different ergodic implications from the neoclassical model. The class of models they were referring to was developed by Azariadis and Drazen (1990). Azariadis and Drazen extended the spillover models of Lucas (1988) and Romer (1986) and showed that, if (local) nonconvexities in the production function were sufficiently strong, then countries that are similar in all aspects except for initial conditions may nevertheless be organized into multiple growth regimes, each of which corresponds to a different steady state for long-run per capita income.

Bernard and Durlauf showed that the multiple-regimes Azariadis–Drazen model was theoretically consistent with a finding of conditional convergence in the data. Therefore, even in the narrowly restricted sense of countries being structurally similar, the finding of a negative coefficient to initial income in the data was no guarantee that countries would converge to a common steady state. Galor (1997) lent further support to the relevance of the Azariadis–Drazen model by arguing that standard ways of augmenting the traditional Solow model increased the likelihood that the true data-generating process followed a multiple-regimes rather than a single steady-state model. Clearly, evidence of multiple regimes and nonlinearities in growth raises questions about misspecification in empirical studies that assume that all countries follow the same growth process, and casts doubt on inferences and policy recommendations that are drawn from these studies.

The work by Bernard and Durlauf has spurred a large quantity of research searching for the existence of multiple-growth regimes. One direction of this new research has been to argue that the finding of parameter heterogeneity in the neoclassical model may be suggestive of the existence of multiple growth regimes. In a seminal work, Durlauf and Johnson (1995), employing a classification and regression tree methodology, implemented a version of Azariadis and Drazen's model and showed that there was evidence in the data to suggest that countries grouped according to initial per capita income and literacy rates correspond to four different growth regimes. Their work has inspired a long list of confirmatory works using a wide variety of econometric approaches (for example, Bloom et al. 2003; Canova 2004; Durlauf et al. 2001; Kourtellos 2005; Liu and Stengos 1999; and Tan 2005).

While there now is a strong consensus in the literature that there exists substantial heterogeneity

across countries, it should be emphasized that this finding is only suggestive of multiple-growth regimes and is not conclusive evidence of it. These heterogeneities could arise because of small deviations in the specification of the production function (see Masanjala and Papageorgiou 2004) which need not correspond to multiple-growth regimes. Further, even within the context of Azariadis–Drazen model, if non-convexities in the production function are not strong enough, the finding of parameter heterogeneity would not imply the existence of multiple regimes (see Durlauf and Johnson 1995, Figure 2).

An alternative approach to investigating the existence of multiple regimes or convergence clubs has focused on the evolution of the world distribution of per capita income. The aim of this research has been to look for evidence of emerging multimodality (typically, bimodality) in the world income distribution. A secondary aim has been to evaluate the degree of churning within the multimodal distribution. If the world income distribution is characterized by emerging multimodality with little evidence of countries moving freely within the distribution (that is, churning), then this finding would suggest, in a manner analogous with the finding of multiple-growth regimes, that global income inequality is real, intensifying and persistent in nature. In fact, these are the precise findings by Quah (1993). By estimating transition probabilities for the cross-country per capita income distribution, Quah finds emerging 'twin peaks' in the world income distribution as well as substantial persistence within the distribution. Quah's seminal work has been confirmed by subsequent work (for example, Bianchi 1997; Fiaschi and Lavezzi 2003; and Paap and van Dijk 1998) even though there had been questions about the robustness of his initial methodology (see Kremer et al. 2001).

While the findings of the 'twin peaks' literature have been suggestive of growth nonlinearities and multiple equilibria, it is not definitive. It is quite possible, for instance, that the aggregate production functions across countries actually exhibit decreasing marginal productivity of capital, so that there is only one steady state. However, other growth factors are sufficiently strong to overcome the convergence effect of diminishing marginal returns to produce divergence and bimodality in cross-country incomes nevertheless. Without an explicit theory to explain the observed income divergence, there is also the question of whether the bimodality in the cross-country income distribution is a transitional or permanent feature of growth (see Galor 1997; Lucas 2000).

Conclusion

Nonlinearities in growth have been highly influential in shaping the thinking of both growth theorists and empiricists in recent years. The work on multiple-growth regimes and the world income distribution suggests that there may exist growth factors strong enough to overcome the decreasing marginal productivity of the neoclassical production function, thereby producing increasing inequality across countries. Nevertheless, while an increasingly large body of work finds evidence that is suggestive of growth nonlinearities, many questions remain open and are the subject of current research. What are the factors that are responsible for generating multiple growth regimes or convergence clubs? Are the effects of these factors transient or permanent? If the former, what are the applicable timescales? This area of research continues to be promising and fruitful.

See Also

- ▶ Balanced Growth
- ▶ Diffusion of Technology
- ▶ Economic Growth, Empirical Regularities in
- ▶ Economic Growth in the Very Long Run
- ▶ Endogenous Growth Theory
- ▶ Growth Take-offs
- ▶ Rosenstein-Rodan, Paul Narcyz (1902–1985)
- ▶ Structural Change

Bibliography

Azariadis, C., and A. Drazen. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105: 501–526.

Barro, R.J., and X. Sala-i-Martin. 1992. Convergence. *Journal of Political Economy* 100: 223–251.
Basu, S., and D.N. Weil. 1998. Appropriate technology and growth. *Quarterly Journal of Economics* 113: 1025–1054.
Bernard, A., and S. Durlauf. 1996. Interpreting tests of the convergence hypothesis. *Journal of Econometrics* 71: 161–173.
Bianchi, M. 1997. Testing for convergence: Evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* 12: 393–409.
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.
Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.
Durlauf, S., and P. Johnson. 1995. Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics* 10: 363–384.
Durlauf, S., A. Kourtellos, and A. Minkin. 2001. The local Solow growth model. *European Economic Review* 45: 928–940.
Fiaschi, D., and M. Lavezzi. 2003. Distribution dynamics and nonlinear growth. *Journal of Economic Growth* 8: 355–378.
Fleming, J.M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65: 241–256.
Galor, O. 1997. Convergence? Inferences from theoretical models. *Economic Journal* 106: 1056–1069.
Galor, O., and D.N. Weil. 2000. Population, technology, and growth: From the Malthusian regime to the demographic transtition and beyond. *American Economic Review* 90: 806–828.
Howitt, P., and D. Mayer-Foulkes. 2005. R&D, implementation, and stagnation: A Schumpeterian theory of convergence clubs. *Journal of Money, Credit and Banking* 37: 147–177.
Koopmans, T.C. 1965. On the concept of optimal growth. In *The econometric approach to development planning*. Amsterdam: North Holland.
Kremer, M., A. Onatski, and J. Stock. 2001. Searching for prosperity. *Carnegie-Rochester Conference Series on Public Policy* 55: 275–303.
Kourtellos, A. 2005. Modeling parameter heterogeneity in cross-country growth regression models. Working paper, Department of University of Cyprus.
Lewis, A. 1956. *The theory of economic growth*. London: Allen & Unwin.
Liu, Z., and T. Stengos. 1999. Non-linearities in cross-country growth regressions: A semiparametric approach. *Journal of Applied Econometrics* 14: 527–538.
Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.
Lucas, R. 2000. Some macroeconomics for the 21st century. *Journal of Economic Perspectives* 11: 159–168.
Maddison, A. 2004. *The world economy: Historical statistics*. Paris: Development Studies Centre, OECD.
Mankiw, N.G., D. Romer, and D. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–437.
Masanjala, W., and C. Papageorgiou. 2004. The Solow model with CES technology: Nonlinearities and parameter heterogeneity. *Journal of Applied Econometrics* 19: 171–202.
Murphy, K., A. Shleifer, and R. Vishny. 1989. Industrialization and the big push. *Journal of Political Economy* 97: 1003–1026.
Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. New York: Oxford University Press.
Paap, R., and H. van Dijk. 1998. Distribution and mobility of wealth of nations. *European Economic Review* 42: 1269–1293.
Parente, S.L., and E.C. Prescott. 1994. Barriers to technology adoption and development. *Journal of Political Economy* 102: 298–321.
Quah, D.T. 1993. Empirical cross-section dynamics for economic growth. *European Economic Review* 37: 426–434.
Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037.
Rosenstein-Rodan, P. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53: 202–211.
Rostow, W.W. 1960. *The stages of economic growth*. Oxford: Oxford University Press.
Sala-i-Martin, X. 1996. The classical approach to convergence analysis. *Economic Journal* 106: 1019–1036.
Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62: 143–151.
Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70(1): 65–94.
Swan, T.W. 1956. Economic growth and capital accumulation. *The Economic Record* 32: 334–361.
Tan, C.M. 2005. *No one true path to development: Uncovering the interplay between geography, institutions, and ethnic fractionalization in economic development*. Mimeo: Tufts University.

Economic Growth, Empirical Regularities in

Steven N. Durlauf and Paul A. Johnson

Abstract

The evolution of economic growth theory throughout the post-war period has been deeply influenced by the effort to explain

broad patterns in cross-country behaviour. We discuss some of the salient empirical regularities associated with neoclassical and new growth economics and consider the shift in focus that has occurred. We first describe the stylized facts of Kaldor that played an important role in the assessment of neoclassical growth models. Next, we consider how a switch in focus to a different class of regularities is associated with the new growth economics that began in the 1980s and dominates contemporary research.

Keywords

Economic growth; Elasticity of substitution; Endogenous growth; Factor shares in national income; Kaldor, N; Labour productivity; Labour's share of income; Model uncertainty; Neoclassical growth theory

JEL Classification

O47

The evolution of economic growth theory throughout the post-war period has been deeply influenced by the effort to explain broad patterns in cross-country behaviour. In this entry, we discuss some of the salient empirical regularities associated with neoclassical and new growth economics and consider the shift in focus that has occurred. We first describe the role of empirical regularities in neoclassical growth theory as it emerged in the 1950s. Next, we consider how a switch in focus to a different class of regularities is associated with the new growth economics that developed in the 1980s and continues to dominate contemporary research. Finally, we assess this shift. Durlauf et al. (2005) contains details of the data and methods used to substantiate the claims made here.

Empirical Regularities and Neoclassical Growth

Neoclassical growth theory is commonly associated with Kaldor's (1961) well-known 'stylized facts' of long-run economic behaviour, which primarily focused on the invariance of long run behaviour for advanced economies. Four of his six facts – (1) the constancy of the growth rate of output per worker over long time horizons, (2) the constancy of the growth rate of capital which is lower than the growth rate of the labour supply, (3) the absence of any systematic trends in the capital–output ratio and (4) the constancy of the rate of profit (and, by implication with the other facts, factor shares in national income) – emphasize common behaviour across countries. Only the fifth and sixth facts – the presence of substantial differences in output per worker across countries, and the positive relationship between the rate of profit and the investment – output ratio – focus on heterogeneity. Kaldor (1957) cites the prediction of constant factor shares as an important test of alternative growth models. An important empirical study at the time was Klein and Kosobud (1961) who investigated constancy by testing for a trend in labour's share, finding none using US data from 1900 to 1953.

While these facts are generally cited as a motivation of neoclassical growth models, their actual relationship to the theory is in fact more complicated. In Solow (1956), for example, the objective is the explanation of long-run economic growth and the constancy of factor shares is only mentioned in passing as an implication of the Cobb–Douglas technology. Indeed Solow (1958) criticizes the literature studying the constancy of factor shares for lacking a precise notion of constancy given that exact constancy cannot reasonably be expected. Bronfenbrenner (1960) argues that, for a wide range of values of the elasticity of substitution between capital and labour, and for reasonable variation in the capital–labour ratio, the theoretical variation in factor shares is consistent with that observed. He concludes that the constancy or otherwise of factor shares is not useful in the assessment of (distribution) theories. Put differently, the first three of Kaldor's stylized facts seem most important to understanding the motivation of the neoclassical program; Solow (2000, p. 4) (in a discussion originally published in 1970) remarks that growth theory is largely

> devoted to analyzing the properties of steady states and to finding out whether an economy not initially in a steady state will evolve into one . . .

How do Kaldor's stylized facts appear from the vantage point of modern empirical growth research? Barro and Sala-i-Martin (2004, pp. 12–16) assess the concordance of Kaldor's stylized facts with the data and conclude that, with the exception of the constant rate of profit, each of the first five holds 'reasonably well' for developed economies. They cite evidence suggesting some tendency for the real rate of return to decline in some economies. The evidence they present, and that which we discuss below, shows that, at least as far as it concerns the rate of growth of labour productivity, the sixth of Kaldor's facts also fits well with the data.

Kaldor's stylized facts are therefore of contemporary use in understanding long- run output behaviour. That said, the facts are no longer central to the research efforts in growth economics as other regularities (or the lack thereof) have become the primary focus of research. We therefore turn to those regularities that have become the focus of contemporary work.

Empirical Regularities and the New Growth Economics

The renaissance in growth theory associated with the rise of endogenous growth models was influenced by interest in the determinants of heterogeneity in growth experiences. While not usually called stylized facts, there is a set of general propositions about heterogeneity that have been very important in influencing research. The most prominent global features evident in the data are the divergence in living standards over the past three centuries and the large disparities in living standards at the end of the twentieth century. By modern standards, all countries were poor in 1700 but since then sustained growth, first in the United Kingdom and parts of Western Europe, and more recently in the United States and parts of the Asia–Pacific region, has resulted in large cross-country differences in living standards. In 2000 average GDP per worker in some countries was about one-fiftieth that in the United States while more than 40% of the world's population lived in countries with average levels of GDP per worker of no more than ten per cent of that in the United States.

Divergence in living standards over the 1960–2000 period is also evident in the large group of countries covered by the Penn World Tables (PWT) (Heston et al. 2002). While a substantial group of countries has exhibited prolonged growth over this period, there remains a large mass of countries at the bottom of the distribution. One result was a hollowing out of the middle of the distribution – a phenomenon labelled 'twin peaks' by Quah (1996; 1997). Moreover, there is strong persistence within the cross-country income distribution with a Spearman rank correlation of 0.84 between GDP per worker in 1960 and that in 2000. This degree of correlation is not peculiar to the PWT data. Easterly et al. (1993) report a rank correlation of 0.82 between GDP per capita in 1988 and that in 1870 for the 28 countries in Maddison (1989). This sense of a lack of mobility is reinforced by Bianchi (1997), who found that very few of the possible crossings from one end of the distribution to the other actually occurred between 1970 and 1989.

The persistence in levels of GDP per worker contrasts sharply with the wide cross-country variation in the growth rates of GDP per worker especially for those countries with relatively low levels of GDP per worker in 1960. The data show scant support for the proposition that the countries of the world are converging to a common level of income per person or for the belief that poor countries have always grown slowly. Both growth 'miracles' – countries exhibited consistently strong growth over the 1960–2000 period – and growth 'disasters' – countries that did poorly, often having negative average growth rates – are present in the data. East and South East Asian countries are well represented among the former group while the later is dominated by countries in sub-Saharan Africa. Taiwan, for example, grew at an average annual rate of over six per cent during this 40-year period and increased GDP per worker by a factor of 11 in the process. Hong Kong,

Korea and Singapore were not far behind in either respect. By contrast, Mauritania, Senegal, Chad, Mozambique, Madagascar, Zambia, Mali, Niger, Nigeria, the Central African Republic, Angola and the Democratic Republic of the Congo all had negative average growth over this period.

For most countries, the average growth rate from 1980 to 2000 was lower than that from 1960 to 1980. The notable exceptions to this observation are China and India. Moreover, past growth does not seem to be a good predictor of future growth as, for example, the correlation between growth in 1960–1980 and that in 1980–2000 is just 0.40. Easterly et al. (1993) suggest that the lack of persistence in growth rates indicates the importance of good luck in economic development. Nevertheless, the cross-decade correlations in growth rates have tended to increase during the 1960–1980 period, indicating a sorting of countries into distinct groups of winners and losers.

There seems to be little relationship between the 1960 level of GDP per worker and subsequent average growth rates. The cross-country dispersion of growth rates tends to fall as initial income rises largely due to the rarity of poor performance among the countries with relatively high levels of GDP per worker in 1960. There is, however, a close relationship between geographical group membership and economic growth between 1960 and 2000. As alluded to above, the countries of sub-Saharan Africa performed poorly over this period, with three-quarters of them growing at an average annual rate of less than just 1.3%. The countries in South and Central America did somewhat better with three-quarters of them having grown at an average of less than 1.5%. Among the East and South East Asian countries, three-quarters grew at an average rate of over 3.8%, and a similar fraction of the South Asian countries grew at over 1.9%.

Many of the poor countries of the world were unable to break out of stagnation between 1960 and 2000. A country growing at two per cent per year for 40 years would enjoy a 120% increase in income per person over that period. Yet, between 1960 and 2000, about a quarter of countries never exceeded their 1960 income level by more than 60%, and about ten per cent of countries never exceeded their 1960 level by more than 30%. One reason for this stagnation is the disposition of some economies to large, abrupt output collapses. About half of countries experienced a 3-year output collapse of 15% or more between 1960 and 2000. Over the same period, the largest 3-year output collapse in the United States was 5.4%, and in the United Kingdom 3.6%, both in 1979–1982.

In sum, there are large cross-countries disparities in GDP per worker and hence in living standards. These disparities have grown wider since 1960 and the middle of cross-country income distribution has thinned since 1960. There is substantial immobility in a country's position in the distribution. Growth rates are much less persistent and have tended to fall since 1980. In general, the countries of sub- Saharan Africa performed poorly over the 1960–2000 period. The countries in South and Central America did somewhat better while the South Asian countries did better still. The East and South East Asian countries did best of all.

The Changing Empirical Focus of Growth Economics

The two sets of empirical regularities we have described, while appearing to differ greatly in terms of their implications for understanding the determinants of the growth process, may in fact be reconciled. A key difference between neoclassical and modern growth economics is its domain of explanation: whereas neoclassical theory attempted to understand the long-run behaviour of advanced industrialized economies, the new growth economics attempts to understand worldwide growth patterns. As a result, the differences between the advanced industrialized economies and the rest of the world take on primary importance. Lucas (2002, pp. 2–3) describes his motivation as

> to see whether modern growth theory could also be adapted for use as a theory of economic development. Adaptation of some kind was evidently necessary: The balanced path of growth theory, with

> constant income growth, and the assumed absence of population pressures, obviously did not fit all of economic history or even all the behavior that can be seen in today's world. The theory is, and was designed to be, a model of the recent past of a subset of countries.

Thus, as the domain of inquiry in growth economics has evolved, the stylized facts of interest have shifted to identifying features of international divergence rather than international convergence.

Further, the effort to identify patterns that characterize the differences in crosscountry growth experiences has led to empirical research that focuses on the identification of particular factors in generating the divergence. Theoretical work in growth economics moved away from the traditional emphasis on factor accumulation and towards the analysis of a wide range of social, historical, geographic, and political factors as sources of cross-country heterogeneity. For example, a major strand of contemporary research focuses on the ways that institutional quality affects growth and development; see Acemoglu et al. (2005) for a detailed survey. The richness of the modern growth literature has led to the widespread use of regression methods to allow for the simultaneous consideration of multiple growth determinants, with a focus on identifying which determinants in fact matter.

The move towards regression methods as the basis for empirical growth research has altered the nature of the sorts of regularities that link data and theory. It is still the case that theoretical analyses are often motivated by the identification of a bivariate relationship between some factor of interest and growth rates. However, relationships of this type do not represent basic growth regularities in the way that Kaldor's stylized facts did. The reason for this transition is that the different growth factors that have expanded the domain of growth economics are typically mutually consistent (Brock and Durlauf 2001) and so the empirical significance of one factor can only be assessed when others are considered as well. Put differently, the finding of a bivariate relationship, or lack thereof, can always be rationalized as reflecting a failure to control for other factors.

As a result, the empirical regularities that matter for contemporary research, such as the coefficient relating a measure of institutional quality to growth, are derivative from statistical analyses of the entire growth process. But statistical models of growth are subject to many forms of model uncertainty, ranging from uncertainty about the appropriate theories to employ to uncertainty about the empirical measurement of the qualitative factors identified by a theory to uncertainty about the details of the statistical specification of a model; see Durlauf and Quah (1999) and Durlauf et al. (2005) for a delineation of these issues. Model uncertainty has meant that there is relatively little consensus on the empirically salient determinants of growth and so little consensus on which regularities should be of primary interest. Thus current growth economics has been handicapped as different papers identify different salient empirical regularities, with inadequate attention to the robustness of such claims. The development of sturdy inferences about the growth process thus represents a very active area of current work.

See Also

- ▶ Economic Growth
- ▶ Endogenous Growth Theory
- ▶ Growth Accounting
- ▶ Level Accounting

Bibliography

Acemoglu, D., S. Johnson, and J. Robinson. 2005. Institutions as the fundamental cause of long-run growth. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Barro, R., and X. Sala-i-Martin. 2004. *Economic growth*, 2nd ed. Cambridge, MA/London: MIT Press.

Bianchi, M. 1997. Testing for convergence: Evidence from nonparametric multimodality tests. *Journal of Applied Econometrics* 12: 393–409.

Brock, W., and S. Durlauf. 2001. Growth empirics and reality. *World Bank Economic Review* 15: 229–72.

Bronfenbrenner, M. 1960. A note on relative shares and the elasticity of substitution. *Journal of Political Economy* 68: 284–7.

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.
Durlauf, S., P. Johnson, and J. Temple. 2005. Growth econometrics. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.
Easterly, W., M. Kremer, L. Pritchett, and L. Summers. 1993. Good policy or good luck? Country growth performance and temporary shocks. *Journal of Monetary Economics* 32: 459–83.
Heston, A., R. Summers, and B. Aten. 2002. *Penn world table version 6.1*. Philadelphia: Center for International Comparisons at the University of Pennsylvania (CICUP).
Kaldor, N. 1957. A model of economic growth. *Economic Journal* 67: 591–624.
Kaldor, N. 1961. Capital accumulation and economic growth. In *The theory of capital, Proceedings of a Conference held by the International Economic Association*, ed. F. Lutz and D. Hague. London: Macmillan.
Klein, L., and R. Kosobud. 1961. Some econometrics of growth: Great ratios of economics. *Quarterly Journal of Economics* 125: 173–98.
Lucas, R. 2002. *Lectures on economic growth*. Cambridge, MA: Harvard University Press.
Maddison, A. 1989. *The world economy in the 20th Century*. Paris: OECD.
Quah, D. 1996. Twin peaks: Growth and convergence in models of distribution dynamics. *Economic Journal* 106: 1045–55.
Quah, D. 1997. Empirics for growth and distribution: Stratification, polarization, and convergence clubs. *Journal of Economic Growth* 2(1): 27–59.
Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Solow, R. 1958. A skeptical note on the constancy of relative shares. *American Economic Review* 48: 618–31.
Solow, R. 2000. *Growth theory: An exposition*. Oxford: Oxford University Press.

Economic Harmony

Israel M. Kirzner

This term has been introduced frequently into economic discussion, and especially into discussions concerning the history of economic thought. Yet there seems to be a good deal of ambiguity as to what it is to mean. Moreover, there has developed considerable disagreement concerning the centrality of the 'harmony' idea to the development of economic thought, and similar disagreement concerning the extent to which the classical economists, in particular, are to be seen as harmony-theorists. We will return a little later to distinguish various different senses that have been attached to the term 'harmony' in economics. For each of these different senses, however, acceptance of the harmony thesis has been held to imply a favourable stance towards a policy of laissez-faire. It is thus not surprising that 18th-century precursors of the notion of harmony have been discovered in Cantillon and in Quesnay (Schumpeter 1954, p. 234). And we are not surprised to find some writers emphasizing the harmony ideas they see in the classical economists, especially in Adam Smith (Halévy 1901–4, p. 89; Heimann 1945, p. 65), while others vehemently question the unqualified identification of these writers with harmony theories (Robbins 1952, pp. 22–9; Samuels 1966, pp. 6–8; Sowell 1974, pp. 16f). It was in the middle of the 19th century that the best–known writings appeared concerning economic harmony. The term appeared in the title of two books by the American economist Henry C. Carey (Carey 1836, 1852). These works were followed by a general treatise stressing the same theme (Carey 1858–60). The term also appeared in the title of a book by the French economic writer Frédéric Bastiat (1850). For a (muted) defence of Bastiat against widespread 19th-century charges that his work in this respect was a crude plagiarism of Carey, see Teilhac (1936, pp. 100–113), who points to the inspiration that both Carey and Bastiat received from J.B. Say. Subsequent references to harmony theories in economics generally tended to be critical, as economists began to argue (from the latter decades of the 19th century into the 20th century) for greater state intervention in market economies on perceived grounds of economic efficiency or economic justice. During most of the 20th century economists, even when they have defended the efficiency and justice of markets, have generally not couched their arguments explicitly in terms of harmony theory. Even Ludwig von Mises who, as we shall see, was an important exception to this last generalization, relegated the notion of

harmony to a distinctly subsidiary role in his system. Recent re-awakened attention to 18th-century theories of spontaneous order, especially as rediscovered and expanded in the work of Hayek, has not had the effect of reintroducing the term 'economic harmony' to current usage. We turn now to take notice of the several different (although certainly interrelated) senses in which this term has been used during the history of economics.

Harmony as Flowing from Divine Providence

A harmony 'theory' is not, in this sense, one that flows out of economic science; rather it represents an attitude of (usually religious) optimism and faith, which itself suggests and guides the course of scientific investigation.

Just as Kepler was inspired by the doctrine of harmony in the spheres to discover the laws which govern the orbits of the planets, so the early economists were inspired by the doctrine that there is a harmony of interests in a society to formulate economic laws (Streeten 1954, p. 208).

It was from this sense of the term that Lord Robbins vigorously dissociated the classical school. It was this optimistic doctrine that came to be referred to contemptuously by the German term 'Harmonielehre'. Archbishop Whately, who in 1832 set up a chair of political economy at Trinity College, Dublin, was an influential harmony theorist in this sense. He saw the purpose of the chair as that of combatting the irreligious implications, as he saw them, of Ricardian economics. The early Dublin professors 'were under pressure to present an optimistic or harmonious picture of how the market economy operates' and the resulting critical attitude towards Ricardian theory reflected 'these extrascientific concerns' (Moss 1976, p. 153). A variant of this approach to the harmony doctrine was the Enlightenment view, in which Deistic philosophy perceived a natural order as responsible for 'predetermined harmony' (Mises 1949, p. 239; Heimann 1945, p. 49).

Harmony Theory as the Doctrine of Maximum Satisfaction

When major neoclassical economists such as Marshall (1920, p. 470) and Wicksell (1901, p. 73) referred to harmony theorists, they evidently had in mind those who believed that economic theory demonstrates that free competitive markets generate maximum total satisfaction for society as a whole. 'Harmony theory' thus referred to a very specific conclusion of economic science, a conclusion central to welfare economics, but a conclusion whose validity both Marshall and Wicksell were concerned to refute. Of special concern, in this context, was the issue of whether the new marginal utility doctrines had been successfully deployed by Jevons, or by Walras, to arrive at 'harmony' conclusions similar to those that had been reached, on other grounds, by Bastiat.

Parallel to this sense of harmony was that which attributed *ethical* virtues to the distributive results of competitive markets. Thus J.B. Clark's demonstration of the justice of marginal-productivity incomes is seen as 'harmony doctrine' (Myrdal 1932, p. 148).

Harmony Doctrine as the Denial of Class Conflict

One sense in which harmony doctrines have been understood throughout the history of economics is that in which it is sought to demonstrate the mutual compatibility of the interests of the various individuals and groups in society. In particular, such doctrines tend to dismiss the notion of inherent class conflict under capitalism. A 20th-century economist who has himself emphasized this idea of harmony of interests in the market society, put the genesis of this idea as follows:

When the classical economists [asserted 'the theorem of the harmony of the rightly understood interests of all members of the market society' they were stressing] two points: First, that everybody is interested in the preservation of the social division of labour, the system that multiplies the productivity of human efforts. Second, that in the

market society consumers' demand ultimately directs all production activities (Mises 1949, p. 674).

Mises, indeed, saw these ideas as important results of economic science, having wide application. 'There is no conflict between the interests of the buyers and those of the sellers, between the interest of the producers and those of the consumers' (Mises 1949, p. 357). Only in the special case of resource monopoly ownership may it happen that the 'emergence of monopoly prices... creates a discrepancy between the interests of the monopolist and those of the consumers' (Mises 1949, p. 680).

Harmony and the Spontaneous Order Tradition

Since the early 1940s F.A. Hayek has succeeded in drawing the attention of economists and others to a line of social analysis since the 18th century, an approach often termed the 'spontaneous order tradition'. The emphasis, in this tradition, is on the evolution of institutions and social outcomes 'which are indeed the results of human action, but not the execution of any human design' (Ferguson 1767, p. 187, cited in Hayek 1967, p. 96). There is no doubt that the term 'economic harmony' has often been applied as an expression of belief in the *possibility and social benignity of undesigned social outcomes*. To some extent, of course, this sense of the term overlaps those listed above, but the emphasis here is not in the denial of conflict, not on any particular welfare theorem, certainly not on any religiously based optimism, but on the counter-intuitive possibility of orderly results emerging without deliberate design from the spontaneous interplay of independently acting individuals. 'Order' in this context has come to mean 'mutually reinforcing expectations'. The following reference to this notion of harmony expresses this usage of the term:

The great general rule governing human action at the beginning, namely that it must conform to fair expectations, is still the scientific rule. All the forms of conduct complying with this rule are consistent with each other and become the recognized customs. The body of custom therefore tends to become a harmonious system (Carter 1907, p. 331, cited in Hayek 1973, p. 169).

The above survey has been confined to notions of economic harmony believed to be achieved spontaneously, 'naturally', without design. For the sake of completeness it should perhaps be noted that the term 'harmony' has occasionally been used to describe the objective of *deliberate* social policy. Thus a well-known debate was initiated by E. Halévy in his claim that Bentham and the philosophical Radicals subscribed to two partly contradictory principles: the 'economic' principle of 'natural identity' (i.e. harmony) of interests, and the 'juristic' principle of the 'artificial identification of interests' (Halévy 1901–4, pp. 15, 17, 489). Lord Robbins, in disputing Halévy concerning any contradiction in the Benthamite position, refers to the juristic principle as contending it to be 'the function of the legislator to bring about an artificial harmonization of interest' (Robbins 1952, pp. 190f). While occasional references may be found to harmony sought to be artificially accomplished, the term has, in general, been associated almost invariably with harmony achieved undeliberately in a decentralized system.

See Also

- ▶ Bastiat, Claude Frédéric (1801–1850)
- ▶ Enlightenment, Scottish

References

Bastiat, F. 1850. *Les harmonies économiques*. Paris: Guillaumin.

Carey, H.C. 1836. *The harmony of nature*. Philadelphia: Carey, Lea & Blanchard.

Carey, H.C. 1852. *The harmony of interests, agricultural, manufacturing, and commercial*, 2nd ed. New York: Myron Finch.

Carey, H.C. 1858–60. *Principles of social science*. Philadelphia: J.B. Lippincott.

Carter, J.C. 1907. *Law, its origin, growth and function*. New York/London: G.P. Putnam's Sons.

Ferguson, A. 1767. *An essay on the history of civil society*. London.

Halévy, E. 1901–4. *The growth of philosophic radicalism*. Translated from the French by M. Morris, 1928. Boston: Beacon, 1955.

Hayek, F.A. 1967. *Studies in philosophy, politics and economics*. Chicago: University of Chicago Press.
Hayek, F.A. 1973. *Law, legislation and liberty*, Rules and Order, vol. I. Chicago: University of Chicago Press.
Heimann, E. 1945. *History of economic doctrines. An introduction to economic theory*. New York: Oxford University Press.
Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan, 1936.
Moss, L.S. 1976. *Mountifort Longfield: Ireland's first professor of political economy*. Ottowa: Green Hill.
Myrdal, G. 1932. *The political element in the development of economic theory*. Translated from the German by P. Streeten. Cambridge, MA: Harvard University Press, 1954.
Robbins, L. 1952. *The theory of economic policy in English classical political economy*. London: Macmillan, 1965.
Samuels, W.J. 1966. *The classical theory of economic policy*. Cleveland/New York: World.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Sowell, T. 1974. *Classical economics reconsidered*. Princeton: Princeton University Press.
Streeten, P. 1954. Recent controversies. Appendix to Myrdal (1932).
Teilhac, E. 1936. *Pioneers of American economic thought in the nineteenth century*. Translated from the French by E.A.J. Johnson (1936), reprinted, New York: Russell and Russell, 1967
von Mises, L. 1949. *Human action: A treatise on economics*, 3rd ed. Chicago: Regnery, 1966.
Wicksell, K. 1901. *Lectures on political economy,* Vol. I. Translated from the Swedish by E. Classen. London: Routledge and Kegan Paul, 1934.

Economic History

Alexander J. Field

Abstract

Economic history focuses on the historical study of growth and development. Originating in the German historical school and studies of the Industrial Revolution in England, it became professionally differentiated from economics proper with the establishment of associations in Britain (1926) and the United States (1941). As economics continued on its increasingly mathematical and ahistorical path in the 1960s, the 'new economic history' advocated applying theory to history. But its emphasis on data analysis retained a bridge to older traditions. As economists have rediscovered an interest in long-term economic growth, often applying traditional institutional approaches, there is continuing evidence of rapprochement.

Keywords

American Economics Association; Anthropometric history; Cliometrics; Economic development; Economic growth; Economic history; Economic History Association in the United States; Historical School (German); Great Depression; Industrial Revolution; Kuznets, S.; Marshall, G.; Marx, K.; Mathematical models; Mercantilism; New economic history; North, D.; Physiocrats; Royal Economic Society; Smith, A.; Total factor productivity; Weber, M

JEL Classifications

N0

Economic history is a sub-discipline within economics and, to a lesser degree, within history, whose main focus is the study of economic growth and development over time. It is to be distinguished from the history of economic thought, a branch of intellectual history.

Studies in economic growth, whether historical or contemporary, develop and analyse quantitative measures of increases in output and output per capita, emphasizing in particular changes in saving rates and rates of technological innovation and their consequences. Economic development is a larger and more encompassing rubric, also including consideration of the role of cultural changes and changes in formal institutions.

Economic history has its origins in two main traditions. The first is the German historical school, a group of scholars in the 19th and early 20th centuries, including Gustav Schmoller and Max Weber, who ranged widely over human history with special emphasis on the consequences of institutional variation for economic as well as political performance. The second tradition stems from the efforts of a group of writers who

viewed the complex of innovations in steam power, iron manufacture, and textiles in late 18th-century Britain as an epochal event – an industrial revolution – equivalent in its significance for human welfare to the Neolithic revolution which gave birth to agriculture around ten millennia earlier. The study of the causes, dimensions and consequences of the emergence of sustained increases in per capita incomes – what Simon Kuznets (1966) called modern economic growth – along with a focus on the consequences of institutional variation, continues to define much of what economic historians do.

Although historians have practised their craft at least since the time of Herodotus (the fifth century BC), economics emerged as a separate social science with the work of Adam Smith or, perhaps, as some have argued, that of the Mercantilists and the Physiocrats. Classical economists, with the notable exception of Ricardo, were almost all also historical economists. The reader of Smith, Mill, Marx, or even Marshall ploughs through thick volumes in which propositions in economic theory are embedded in often lengthy descriptions of historical events or the course of economic history. Throughout most of the 19th century, the divide between economists and economic historians was weak.

With the professionalization of economics that picked up speed in the 20th century (the American Economics Association was founded in 1885; the Royal Economic Society in 1890), economic history began to emerge as a distinct and to some degree separate sub-discipline. The Economic History Society was founded in Britain in 1926; the Economic History Association in the United States in 1941. The trend towards a separate identity accelerated in the third quarter of the 20th century, with the increasing emphasis within economics on formal mathematical modelling and the weakening within the general profession of ties to historical traditions. In the economic history societies, in contrast, those trained as historians as well as economists remained active; in Britain, distinctiveness was accentuated by the establishment of separate university departments of economic history.

As the intellectual paths taken by economics and economic history seemed increasingly to diverge, a countervailing intellectual movement known variously as 'cliometrics' or the 'new economic history' emerged. Its pioneers knew their history, but emphasized by argument and example that, if economic history was to remain influential within economics, it had to make more use of formal models as well as place increased emphasis on quantitative (rather than just qualitative) data and more advanced statistical techniques (econometrics) to analyse them.

The use of mathematical models was anathema within historical traditions, but by the 1960s widely accepted in economics. Thus, the new economic history represented something of a gauntlet thrown down to those trained in history or allied with its traditions. The push for quantitative data analysis, in contrast, was more cross-cutting in the challenges it implied. Many traditional economic historians had in fact examined such data, although the statistical techniques they used were often quite rudimentary. Within some economic circles, on the other hand, an emphasis on data was becoming suspect. Here, some scholars were comfortable with the evolution of economic theory as a branch of applied mathematics, constrained and judged by the rules of logic and consistency, but governed in its realism, if at all, by intuition rather than systematic empirical inquiry.

The effort to force formal theory upon traditional economic history often lacked acknowledgement that the relation between economic history and formal theory might usefully be a two-way street. The emphasis on data analysis, in contrast, offered a bridge between economics and economic history. It helped reaffirm within economics the importance of empirical inquiry, and encouraged those historically trained to become more sophisticated in their statistical analyses.

Nevertheless, the stress on quantitative data could not help but draw attention away from economic history's traditional concern with legal and institutional variation, where the source documents were almost uniformly qualitative. How would this theme, one of the defining features of

economic history since its inception, survive the new economic history? The initial 'solution' was to try to make institutions endogenous. Blending a mix of influences from technologically deterministic Marxism to the emerging law and economics and public choice literatures, a number of scholars suggested that institutions could be understood as epiphenomenal: reflective of more fundamental givens. The high point of such efforts was probably the short book by North and Thomas (1973).

These efforts, however, gradually disintegrated under the force of the ad hoc twists required to make the framework consistent with known historical evidence (Field 1981), and even proponents such as North eventually backed away from this agenda. Formal rules often vary where technologies and endowments are similar, and are often similar when more fundamental givens differ, and such variation has consequences for economic performance. Had the endogenization initiative been successful, it would have eliminated from economic history one of the most important perspectives it offers to general economics.

The old economic historians had taken it as obvious that, at critical historical junctures, changes in formal institutions such as laws or constitutions had powerful influences on the course of a country or region's economic development, and that these changes were not always predictable *ex ante.* The breakdown of the former Soviet Empire, and the opportunities afforded to Western scholars actually to influence the design of formal institutions, gave a powerful impetus to returning to thinking about such designs as consequential, and increasingly this perspective came to be reflected in research by scholars who did not necessarily think of themselves as economic historians.

If the main subject of economic history continues to be the history of economic growth and development, the influence of variations in formal and informal institutions in both the private and public arenas will remain an important theme. These institutions and a broader economic culture help structure the environment in which individuals pursue their interests. But the success of an economy in raising output and output per person also depends on available technologies, on the size, composition, and characteristics of the labour force, on natural resources, and on the accumulation of physical capital. The study of the evolution of these inputs suggests some of the other themes around which economic historians organize their work. In particular, there is a rich tradition, particularly in the United States, examining issues in and applying methods from modern labour economics within an historical context.

The basic agenda of economic history has not changed since the first edition of *The New Palgrave.* Interest in the causes and dimensions of the Industrial Revolution, for example, remains strong, particularly in Britain. But the field has evolved in new directions, with several discernable trends. First, scholars have concerned themselves with a broadening range of topics under the umbrella of growth and development. In the 1960s and 1970s, especially in the United States, railroads and slavery dominated much of the discussion. In recent decades, it is not possible to point to one or two issues around which research and discussion has coalesced to the same degree. Instead, there have been a number of new initiatives; one example would be the growing exploitation of anthropometric data to make inferences about variation in standards of living.

Associated with this has been a broadening of the scope of the discipline, both in terms of the countries in which economic history research is conducted and in the geographical range of topics, which extends, somewhat more so than in earlier decades, beyond Western Europe and North America to Asia, Latin America, Australia, and Africa. One illustration of this has been a range of cross-national studies, exploring such issues as economic convergence.

A third trend has been a growing willingness to think of the 20th century as an historical epoch in its own right. When the new economic history began, the Second World War had barely ended and the Great Depression was recent history. The main focus of research was the 18th and 19th centuries. Treating the 20th century as an historical period promises to reduce the gap between economics and economic history. The Great

Depression, of course, continues to attract attention, but interest in the 20th century is beginning to expand beyond this. The data and events of recent decades can now more easily be seen in an historical context. The result can be a smoother continuum between topics understood as economic history and the analysis of contemporary data.

Placing more recent developments within a longer-run perspective has already begun to pay important dividends. Many trends that economists and economic historians expected at mid-century would characterize the 20th century as a whole moderated, became erratic, or in some cases reversed themselves in the last quarter of the century (Field 2001). In 1950, for example, it looked as if the United States (and other countries) would continue to experience decreases in wealth and income inequality, robust and perhaps rising shares of union membership in the labour force, a growing role for government, and a continuing high contribution of total factor productivity (TFP) growth to growth in output per hour. In fact, inequality has generally increased, union membership has fallen, and TFP growth basically disappeared in much of the developed world between 1973 and the 1990s. The size and role of government, which many predicted would continue to expand, has in fact displayed a more complex dynamic.

A fourth and related trend has been a reinvigoration within mainstream economics of interest in what has always been a primary subject of economic history: economic growth. Much of economic theory in the 1950s and 1960s modelled production and allocation within a static economy. The revived interest in the study of growth, combined with the growing willingness of economists to adopt traditional institutional approaches, reflects the persisting influence of the original concerns and approaches of economic history within the larger profession.

Whatever the labels people apply to themselves and others, if we want better understanding of the processes of growth and development, we will continue to need scholars familiar with how to work with data and interpret the influences on economic outcomes of institutional, political, and cultural variation. Doctoral training with a specialization in economic history is well suited to imparting such knowledge and the skills for acquiring it, capabilities that will remain essential in developing improved theory and policy in the area.

See Also

- ▶ Cliometrics
- ▶ Growth and Cycles
- ▶ Growth and Institutions
- ▶ Historical School, German
- ▶ Institutionalism, Old
- ▶ Technical Change

Bibliography

Cairncross, A. 1989. In praise of economic history. *The Economic History Review* 42: 173–185.

Field, A. 1981. The problem with neoclassical institutional economics. *Explorations in Economic History* 18: 174–198.

Field, A. 1987. *The future of economic history*. Boston: Kluwer-Nijhoff.

Field, A. 2001. Not what it used to be: The Cambridge economic history of the United States, vols. II and III. *Journal of Economic History* 61: 806–818.

Kuznets, S. 1966. *Modern economic growth: rate, structure, and spread*. New Haven: Yale University Press.

McCloskey, D. 1987. *Econometric history*. Basingstoke: Macmillan Education.

North, D., and R. Thomas. 1973. *The rise of the Western world*. Cambridge: Cambridge University Press.

Solow, R. 1985. Economic history and economics. *American Economic Review* 75: 328–331.

Economic Impact of the Olympic Games

Andrew Zimbalist

Abstract

The Olympic Games are among the largest and most visible sporting events in the world. Every two years, the world's best athletes

from some 200 countries come together to compete in lavish new venues in front of thousands of spectators. Hundreds of millions of sports fans worldwide watch the Games on television. Although Pierre de Coubertin, who founded the modern Olympics in the late 19th century, may have had altruistic, idealistic notions of pure amateur competition, unsullied by financial motivations, the Olympic Games have become a big business. The participants are effectively professional athletes; the organizers are highly compensated, professional bureaucrats; hosting the Games involves huge construction and renovation projects that take nearly a decade to complete, and these expenditures are usually justified by claims of extraordinary economic benefits that will accrue to the host city or region as a direct result of hosting the Games. This article examines the financing of the Olympic Games, explores how the awarding of the Games has become a high-stakes contest, and analyzes the costs of running the Games and their economic impact on the host city and nation.

Keywords

Economics of sport; International Olympic Committee (IOC); Major events; Multiplier; Olympic bid; Olympic Games; Sport finance; Sport infrastructure

JEL Classifications

L31; L83; L88; R1

Financing the Olympics

The modern Olympic Games began in 1896, but it was not until 1976 that a watershed event shook up the financing model for the Games and set the Olympics on its current economic course. In that year the city of Montreal hosted the Summer Games, which were originally predicted to cost $124 million. In fact, Montreal incurred a debt of $2.8 billion (approximately $10 billion in 2010 dollars) that was only finally paid off in 2005 (Burton 2003). Annual debt service created a large budgetary hole for the city for three decades.

By the end of the Montreal Games, the 1980 Olympics had already been set for Moscow, but no city wanted to bid for the right to host the 1984 Games. After some scrambling, Los Angeles agreed to host the Games, but only on the condition that it took on no financial obligation. With no alternative, the International Olympic Committee (IOC) accepted this condition and Los Angeles was awarded the 1984 Summer Games on 1 July 1978.

That year also marked the first significant relaxation of Olympic amateur rules under then IOC president Lord Killanin. Rule 26 of the Olympic Charter was modified so that athletes were allowed openly to earn money from endorsements, if the money went to their national sports federation or their country's National Olympic Committee (NOC). The receiving organization was then permitted to pay the athlete's expenses, including 'pocket money'. 'Broken-time' payments for time away from the athlete's regular job were also authorized if the athlete had a regular job. But the rule continued to declare that professional athletes were ineligible.

During the 1980–2001 reign of IOC President Juan Antonio Samaranch complete professionalization and commercialization of the Olympics were realized. In 1982, the amateur rules were revised to permit payments into a trust fund that provided expenses during the athlete's active career – and substantial sums thereafter. Eventually, decisions about accepting professionals were left to the International Federation (IF) of each sport. The new professional era was heralded during the 1992 Games in Barcelona, when the USA sent its 'Dream Team' of National Basketball Association (NBA) stars which went on to win the gold medal. Nominally, for the 2004 Games in Athens, boxing was the only sport that did not accept professionals, but even this distinction is dubious, because the National Olympic Committees (NOCs) of many countries gave their boxing medalists cash prizes.

These changes also led to increased commercialization and increased TV and sponsorship money, which in turn led to corruption and

scandal within the IOC. Before 1980, the 112 IOC representatives had to pay their own way to cities bidding for the Games. Within a year they were receiving first-class tickets and all expenses, as well as lavish entertainment. Outrageous tales of the excesses enjoyed by IOC representatives abounded.

Revelations of bribery and corruption around the Salt Lake City bid for the 2002 Games plunged the IOC into scandal: six members were expelled, four resigned and 10 were warned. Since then, the IOC has reformed itself by reducing the number of voters and by officially declaring an outright ban on gifts.

Meanwhile, the modest financial success of the 1984 Games in Los Angeles led to a new era of international competition among cities to host the Games. The relative success of Los Angeles, however, was *sui generis*. Los Angeles had very little construction expense and the chair of the Los Angeles Organizing Committee for the Olympic Games (LAOCOG), Peter Ueberroth, was able to raise substantial sums by selling sponsorships to corporations. LAOCOG generated a modest surplus (just over $300 million) and reset the Olympic financial model for less public and more private financing.

Nonetheless, other host cities have found it impossible to procure the same proportion of private support and have relied upon large public expenditures. Several billion dollars of public monies were committed in Seoul (1988), Barcelona (1992), Nagano (1998), Sydney (2000), Athens (2004) and Beijing (2008). In some cases, the local OCOG ran a modest surplus (20% of any surplus must be shared with the IOC; Preuss 2003, p. 194), but the local government laid out billions of dollars to help finance the activities of the OCOG. In the case of Athens, for instance, the public investment exceeded $10 billion – some of this public investment resulted in improved, more modern infrastructure for the city, but some of it resulted in white elephants. Many facilities built especially for the Games go unused or underutilized after the period of Olympic competition itself, while requiring tens of millions of dollars annually to maintain and occupying increasingly scarce real estate. Public investment for the 2008 Beijing Olympics exceeded $40 billion.

Salt Lake City Olympiad chief and former governor of Massachusetts Mitt Romney questioned whether US cities should enter bids to host the Olympic Games, stating that they were increasingly driven by 'giganticism' with the addition of new sports and more frills.

Present Day Financial Arrangements

The IOC presents the financing of the Olympic Games in terms of the related organizations: the local organizing committee (OCOG), the NOC, the IFs, and itself. The OCOG budget is not the same as the budgetary impact on the local city that hosts the games. The local city and its regional and national government may provide billions of dollars of subsidies to the OCOG, and the OCOG may report a surplus. This surplus has little meaning regarding the budgetary impact on public bodies from hosting the Games.

For instance, for the recent Games hosted in the USA, the federal government provided $1.3 billion in Salt Lake City in 2002; $609 million in Atlanta in 1996; and $75 million in Los Angeles in 1984 (all reckoned in 1999 prices) (Ungar 2000, p. 5). For the 2010 Winter Games in Vancouver, in addition to the provincial government of British Columbia and the federal government of Canada each budgeting $9.1 million to help finance the bidding process, the provincial government was scheduled to put up $1.25 billion to finance the Games (and provide a guarantee to cover cost overruns) and the federal government was budgeted to contribute another $330 million. The city of Vancouver was budgeted to contribute $170.3 million (http://www.mapleleafweb.com/, accessed 22 August 2007). Not surprisingly, financing did not work out as planned, in part due to the worldwide recession of 2008–09. In fact, the IOC provided a $423 million subsidy to the Vancouver Organizing Committee (VANOC) and 'sources said that the IOC agreed to the first-of-its-kind bailout because without it, spending for the games would have come to a screeching halt and major cutbacks would have been

Economic Impact of the Olympic Games, Table 1 Olympic movement revenue (current US $ millions)

Source	1993–96	1997–2000	2001–04	2005–08
Broadcast	1,251	1,845	2,230	2,570
TOP programme	279	579	663	866
Dom. Sponsorship	534	655	796	1,555
Ticketing	451	625	411	274
Licensing	115	66	87	185
Total	2,630	3,770	4,187	5,450

Sources: IOC, *2006 Olympic Marketing Fact File*, http://www.olympic.org/, p. 16; IOC, *2010 Olympic Marketing Fact File*, p. 26; IOC, 2008, *Media Marketing Guide*, p. 4

necessary' (*Sports Business Daily* 2009a). VANOC also received an $87 million public bail-out loan (*Sports Business Journal* 2009a) and Standard & Poor's lowered the city of Vancouver's credit rating due to Olympic financing shortfalls, raising the city's borrowing costs (*Sports Business Daily* 2009b). The *New York Times* reported that: 'The immediate legacy for this city of 580,000 is a nearly $1 billion debt from bailing out the Olympic Village development. Beyond that, people in Vancouver and British Columbia have already seen cuts in services like education, health care and arts financing from their provincial government, which stuck with many other Olympics-related costs' (Austen 2010).

Moreover, it is common practice for the OCOG budget to consist entirely (or almost entirely) of operating, as opposed to capital, expenditures (Preuss 2003, p. 195). Nonetheless, to the extent that the OCOG receives funding from the IOC or from private sources, the lower will be the financing burden that falls on the local, state and national government that hosts the Games. What follows, then, is a discussion of how the IOC distributes the revenue that is collected from the staging of each Olympic Games.

Table 1 presents the total revenue that accrues to the IOC or any of its constituent organizations during each quadrennial Olympic cycle, consisting of one Winter and one Summer Games. It shows a healthy revenue growth in each of the major categories, with television revenues the largest single source of revenue by a factor of 3. The TOP marketing program consists of 11 companies that hold exclusive category sponsorships as the official Olympic company.

TOP (The Olympic Partner) Programme revenues go 50% to the local OCOGs, 40% to the NOCs and 10% to the IOC (International Olympic Committee 2007). Broadcast revenue goes 49% to the host OCOG and 51% to the IOC, which in turn distributes the lion's share of this revenue to the NOCs and IFs. Prior to 2004, the host OCOGs received 60% of broadcast revenue. Beginning in 2012, it has been determined that OCOGs will receive a fixed amount, rather than a fixed percentage, as broadcast revenues continue to rise (Preuss 2003, p. 100). Overall, the IOC has retained 8% of Olympic revenue; the remaining 92% has been shared by the OCOGs, NOCs and IFs.

Table 2 depicts the astronomical growth in television broadcasting revenue for the Summer and Winter Games since 1960. Not surprisingly, the largest share of broadcast revenue comes from the USA. For instance, for the 2004 Athens Games, the IOC contract with NBC yielded $793.5 million, or 53.1% of the total. Following the US rights fee were Europe ($394 million), Japan ($155 million), Australia ($50.5 million), Canada ($37 million) and South Korea ($15.5 million). All told, there were 80 rights holders televising the Athens Games to 220 countries and 2 billion potential viewers worldwide. Ten thousand media personnel were on hand to cover the Games (International Olympic Committee 2007, pp. 51–54).

The US share of total media rights has trended downward over time, from 83.4% during 1986–89, to 60% during 2001–04, to 52.6% during 2009–12 (*Sports Business Journal* 2009b). Because of the high US share, the US Olympic Committee (USOC) has received a

Economic Impact of the Olympic Games, Table 2 Broadcast revenue history (millions, current US $)

Summer		Winter	
Olympic Games	Broadcast revenue	Olympic Games	Broadcast revenue
1960 Rome	1.20	1960 Squaw Valley	0.05
1964 Tokyo	1.58	1964 Innsbruck	0.94
1968 Mexico City	9.75	1968 Grenoble	2.61
1972 Munich	17.79	1972 Sapporo	8.48
1976 Montreal	34.86	1976 Innsbruck	11.63
1980 Moscow	87.98	1980 Lake Placid	20.73
1984 Los Angeles	286.91	1984 Sarajevo	102.68
1988 Seoul	402.60	1988 Calgary	324.90
1992 Barcelona	636.06	1992 Albertville	291.93
1996 Atlanta	898.27	1994 Lillehammer	352.91
2000 Sydney	1,331.55	1998 Nagano	513.49
2004 Athens	1,494.03	2002 Salt Lake	738.00
2008 Beijing	1,739.00	2006 Torino	831.00

Sources: IOC, *2006 Marketing Fact File*, p. 46; IOC, *2010 Market Fact File*, p. 27; IOC, 2008, *Media Marketing Guide*, pp. 4, 6

disproportionate share of the total collected fees. Out of 205 NOCs in 2009, USOC received 12.75% of all media rights fees from the Olympics. (This share had been 10% until 1996, when it was raised.) In 2009, after sharp disagreement, a negotiation between USOC and the IOC led to a new agreement that in 2020 the USOC share would be lowered again. The new level, however, was not agreed upon.

OCOGs do not cover all their expenses from the above sources. For instance, the Nagano OCOG in 1998 had revenues of $990 million, of which approximately $435 million came from the IOC. Similarly, the Salt Lake OCOG had revenues of $1.348 billion, of which approximately $570 million came from the IOC (International Olympic Committee 2007, pp. 82–83; from the Salt Lake Games revenues, the IOC also provided $305 million to the NOCs.).

Economic Results

Economic theory would suggest that any expected local economic benefit would be bid away as cities compete with each other to host the Games. More precisely, with perfect information the city with the highest expected gain could win the Games by bidding $1 more than the expected gain to the second highest city. Such an outcome could yield a small benefit to the winning city, but this would require perfect information and an open market bidding process. In fact, the bidding process is not done in dollar amounts, but comes rather in the form of providing facilities and guaranteeing financing and security. In the post-9/11 world, security costs are far from trivial. Total security costs in Athens in 2004 came to $1.4 billion, with 40,000 security people; Beijing in 2008 was projected to have over 80,000 security personnel working the Games.

It is also widely acknowledged that the bidding process is laden with political considerations. Moreover, the bidding cities are more likely to be motivated by gains to particular private interests within the city (developers, construction companies, hotels, investment bankers, architects, real estate companies, etc.) than by a clear sense that the city as a whole will benefit economically.

In contrast, the IOC views its principal role as promoting sport, not economic development. It requires buildings and infrastructure to be financed with non-Olympic money (Preuss 2003, p. 195).

Accordingly, even though a local OCOG may break even or have a small surplus, the greatest likelihood is that the city itself (and state and national governments) experiences a fiscal deficit

from the Games (of course, not all OCOGs manage to break even: the Albertville OCOG lost $57 million: Burton 2003, p. 3). On the one hand, the only tax revenue that would accrue to host governmental bodies would be from incremental sales and income resulting from hosting the Olympic Games. (Other taxes, such as real estate taxes, might come into play depending on the local tax system and whether or not the Games affected real estate values, positively or negatively. See, for instance, Ahlfeldt and Maennig (2009), which finds a positive impact of sports facilities on real estate values within two miles of a new facility.) The evidence on this score is not encouraging. On the other hand, hosting governmental bodies, together with any private support, must pay for facility construction, upgrade and infrastructural improvements necessitated by the Games. It must also pay for the opening and award ceremonies, transportation of the athletes to the various venues, entertainment, a telecommunications/broadcasting centre, and security, among other things. Naturally, to the extent that some of this public spending is on productive infrastructure, these fiscal deficits may prove to be beneficial in the long term to the economy.

The initial publicized budgets of the OCOGs invariably understate both the ultimate cost to the OCOG and, to a much greater degree, the total cost of staging the Games. The former escalates for several reasons:

1. Construction costs inflate significantly as land values increase with growing scarcity during the roughly ten-year cycle of Olympic host planning, bidding, selection and preparation.
2. It is usually in the interest of the bidding team to under-represent the true costs, as they seek public endorsement.
3. As the would-be host city enters into competition with other bidders, there is a natural tendency to match their competitors' proposals and to embellish their original plans.

The total cost escalates because it includes infrastructure and facility costs, whereas the publicized OCOG budget includes only operating costs. The infrastructure and facility costs usually form the largest component of total expenses, and often do so by a substantial margin.

Thus Athens initially projected that its Games would cost $1.6 billion, but they ended up costing closer to $16 billion (including facility and infrastructure costs). Beijing's projected budget was $1.6 billion, but ended up in total costing over $40 billion (inclusive of facility and infrastructure costs). The 2014 Winter Games in Sochi, Russia, were initially budgeted at around $12 billion; the projected price tag in late 2009 reached $33 billion. Of this, $23 billion would come from public sources (*Sports Business Daily* 2009c).

London expected its 2012 Games to cost under $4 billion, but they are now projected to cost over $19 billion (Carlin 2007; Simon 2006; *Sports Business Daily* 2008a). As expenses have escalated for London, some of the projects have been scaled back, such as the abandonment of the planned roof over the Olympic Stadium. The stadium was originally projected to cost $406 million and will end up costing over $850 million. Further, its construction will be financed by taxpayers and the government has been unsuccessful in its effort to find a soccer or a rugby team to be the facility's anchor tenant after the 2012 Games. This will saddle the British taxpayers with the extra burden of millions of dollars annually to keep the facility operating. It is little wonder that the London Olympics Minister Tessa Jowell stated: 'Had we known what we know now, would we have bid for the Olympics? Almost certainly not'. (*Sports Business Daily* (2008b), citing a story in *Daily Telegraph* (2008). The Olympic Village was to be privately financed, but the plan fell through and will instead cost the taxpayers nearly $1 billion. The government hopes that the apartments will be sold after the Games and the financing will be recouped.)

In a world where total revenue from the Games is in the neighborhood of $4–$5 billion for the Summer Olympics and roughly half that for the Winter Games, costs above these levels mean that someone has to pay. (To be sure, the Winter Games involve fewer participants, fewer venues and less construction; hence the cost of these Games is lower than for the summer Games.) While private companies often contribute a share

of the capital costs (beyond the purchase of sponsorships), host governmental bodies usually pick up a substantial part of the tab. Moreover, as we have seen, not all the money generated at the Games stays in the host city to pay for the Games; rather, close to half the money goes to support the activities of the IFs, the NOCs and the IOC itself.

Thus, while the Sydney OCOG in 2000 reported that it broke even, the Australian state auditor estimated that the Games' true long-term cost was $2.2 billion. In part, this was because it is now costing $30 million a year to operate the 90,000-seat Olympic Stadium. The story was little different for the 2004 Games in Athens, where maintenance costs on the Olympics facilities in 2005 will reportedly come in around $124 million and there appears to be little to no local interest being expressed in the two Olympic soccer stadiums. According to one report, 21 of the 22 stadiums built for the 2004 Summer Games in Athens were unoccupied in 2010.

Similarly, the 1992 Olympics in Barcelona generated a reported surplus of $3 million for the local organizing committee, but it created a debt of $4 billion for the central Spanish government and of $2.1 billion for the city and provincial governments. (The total reported cost of the Barcelona Games was $9.3 billion, of which private sources covered $3.2 billion and public sources covered $6.1 billion (Burton 2003, p. 39). For a related account of large public expenditures on infrastructure for the Salt Lake City Olympics, see Bartlett and Steele (2001), who reported that the US government spent $1.5 billion of taxpayers' money on the purchase of land, road construction, sewers, parking lots, housing, buses, fencing, a light rail system, airport improvements, and security equipment, *inter alia*. Some have argued that a part of these expenditures would have occurred even if Salt Lake City had not hosted the Olympics.) The Nagano Organizing Committee (Winter Games 1998) showed a $28 million surplus, but the various units of Japanese government were left with an $11 billion debt (Burton and O'Reilly 2009).

For all of the foregoing reasons, if there is to be an economic benefit from hosting the Olympic Games, it is unlikely in the extreme to come in the form of improving the budgets of local governments. This raises the question of whether there are broader, longer-term or less tangible economic gains that accrue from hosting the Olympic Games.

How Do the Olympic Games Affect the Economy?

In general, sporting events produce two types of economic benefit: direct and indirect. Direct economic benefits include net spending by tourists who travel from out of town to attend the event; spending on capital and infrastructure construction related to the event; long-run benefits – for example lower transportation costs attributable to an improved road or rail network – generated by this infrastructure; and the effect of hosting a sporting event on local security markets, primarily stock markets. Indirect benefits include possible advertising effects that make the host city or country more visible as a potential tourist destination or business location in the future and increases in civic pride, local sense of community, and the perceived stature of the host city or country relative to other cities or countries.

Among the direct economic benefits generated by the Olympic Games, tourist spending is probably the most prominent. From Table 3, an average of 5.1 million tickets were sold for the past six Summer Olympic Games, including almost six million tickets to the 1984 Games in Los Angeles. The Winter Games are considerably smaller, averaging 1.3 million tickets over the past five Winter Olympics. Even though most spectators buy tickets to multiple events, so that selling five million tickets does not mean that there are five million spectators, and many of the tickets are sold to local residents, especially for the Summer Games, which typically take place in large metropolitan areas, a sporting event of this size and scope has the potential to attract a significant number of visitors from outside the host city. Also, since the Games are often spread over more than two weeks, these visitors may spend a significant amount of time in the host area,

Economic Impact of the Olympic Games, Table 3 Ticket revenues (current $)

Games	Tickets sold (millions)	% capacity	Revenue to OCOG ($million)
1984 Los Angeles	5.7	83	156
1988 Calgary	1.6	78	32
1988 Seoul	3.3	75	36
1992 Albertville	0.9	75	32
1992 Barcelona	3.0	80	79
1994 Lillehammer	1.2	87	26
1996 Atlanta	8.3	82	425
1998 Nagano	1.3	89	74
2000 Sydney	6.7	88	551
2002 Salt Lake	1.5	95	183
2004 Athens	3.8	72	228
2008 Beijing	6.5	96	185

Sources: IOC, *2006 Marketing Fact File*, p. 60; IOC, *2010 Marketing Fact File*, p. 39

generating substantial spending in the lodging, and food and beverage sectors.

The Olympic Games require large spending on constructing and updating venues. In addition to venue construction, hosting the Olympic Games often requires expansive infrastructure to move the participants, officials, and fans to and from the venues. Host cities and regions have also spent considerable sums on roads and airport construction, as well as on the renovation and construction of public transportation systems (Essex and Chalkley 2004). In less developed cities, the building of a modern telecommunications capacity also represents a substantial investment.

After the construction period, Olympics-generated infrastructure can provide the host metropolitan area or region with a continuing stream of economic benefits in the form of reduced production costs and prices charged by local businesses (Rephann and Isserman 1994). The indirect economic benefits generated by the Olympic Games are potentially more important than the direct benefits, and also more difficult to quantify. One possible indirect benefit is the advertising effect of the Olympic Games. Many Olympic host metropolitan areas and regions view the Olympics as a way to raise their profile on the world stage. If hosting the Olympic Games leads tourists who would not have otherwise considered this to be a destination to visit the host city or region, then this advertising effect can generate economic benefits over a long period of time.

The Evidence on Economic Impact

The evidence on the economic impact of the Olympic Games falls into four categories: (1) retrospective evidence based on econometric analysis; (2) case studies of individual Olympic Games; (3) evidence derived from computable general equilibrium (CGE) models; and (4) 'multiplier-based' estimates of future economic impact. Note the important temporal element associated with each of these types of evidence. The 'multiplier-based' estimates are prospective; these studies are basically forecasts of economic benefits that will take place at some time in the future. Because this type of evidence is a forecast, it should be judged by the same criteria as any other economic forecast. The other three types of evidence are retrospective. They are based on an examination of what actually happened in the past when a metropolitan area or region hosted the Olympic Games. This fundamental difference between 'multiplier-based' estimates and other types of evidence is critically important for understanding the differences in estimates of the economic impact of the Olympic Games.

Econometric-based evidence on the economic impact of sporting events uses historic data on the performance of the local economy before, during and after the event takes place. This approach uses statistical methods to determine how much of the past local economic activity could be attributed to the sporting event, and

how much would have taken place without the sporting event occurring.

Case studies of the economic impact use a similar approach to the econometric method. This approach examines past indicators of economic activity, but does not use sophisticated regression techniques. Case studies often examine a broader set of economic indicators than econometric studies, and use unconditional statistical tests like tests of differences in means, cross-tabs, or chi square tests of statistical independence.

CGE models are complex representations of the entire economy, including sectors that are not related to sporting events. These models explicitly account for the interconnected nature of the economy. Because the Olympic Games are large-scale events, involving significant numbers of participants, officials, staff and spectators, the effects of the Games may spread beyond the immediate area and affect a number of distinct sectors of the economy in different ways. CGE models can account for complex economic effects, such as the effect of the additional borrowing needed to finance venue and infrastructure construction on the availability (or price) of funds to finance other construction projects in the economy. CGE models can also explicitly account for long-run economic effects.

The basic idea behind multipliers is straightforward, and emerges from input–output models of the economy. When a consumer purchases a $1 pack of gum at a local store, the economic effects of that transaction extend well beyond the consumer handing a dollar to the cashier, who places that dollar in the till. Some of that $1 in spending finds its way into the pocket of the cashier, in the form of wages; some finds its way into the pocket of the store owner; some into the pocket of the driver who delivered the gum; and so on. If the clerk, store owner and delivery person live in the local community, then this money is further distributed in the local economy as these individuals pay rent, buy groceries, and so on. A multiplier is an analytical device used to estimate the broad economic impact of each dollar spent in the local economy in terms of the total amount of additional revenues earned by firms, the total amount of personal income, and the number of jobs generated in the local economy.

Estimating the economic impact of a sporting event using the multiplier approach is relatively simple in theory. First, estimate the number of people who attend the sporting event; second, estimate the amount of spending by these attendees; third, apply a multiplier to this spending to estimate the broad, overall impact of this spending on the economy. However, on closer examination, this process requires a significant amount of discretionary input on the part of the researcher, and coming up with accurate estimates of several of these components is not a straightforward process.

The Problem with Multipliers

Multiplier-based estimates of the economic impact potentially have several problems; See, for one, the discussion in Crompton (1995). First, multiplier-based estimates stem from the estimate of the number of attendees. New economic impact can only be generated by the spending of spectators, participants and officials from outside the host area. Estimating the total number of attendees is much easier than estimating the number of attendees from outside the host area. From Table 3, 8.3 million tickets were sold to events at the 1996 Atlanta Summer Olympic Games. Some of these tickets were clearly sold to residents of Atlanta. But how many of these 8.3 million tickets were purchased by Atlantans?

Further complicating the process are ‘time switchers’ and ‘casuals’. “Time switchers” are attendees who would have visited the host area at some other time, for some other reason, but instead choose to visit the host area during the sporting event. ‘Casuals’ are attendees who visit the host area at the same time as the sporting event for some other reason and decide to attend the event out of convenience. The spending by both types of attendees needs to be removed from the economic impact estimate, as it cannot be directly attributable to the sporting event. Failure to remove this spending leads to overestimates of the economic impact generated by the event.

Second, multiplier-based economic impact estimates fail to account for crowding out. In many cases, the host area for the Olympic Games is a tourist destination in its own right; tourists would visit this area even if the Olympic Games were held elsewhere; London, for example, is a major tourist destination. Crowding out takes place when outside visitors attending the Olympic Games buy hotel rooms, meals and other travel-related goods and services that would have been purchased by other visitors absent the Olympic Games. Crowding out implies that each dollar of new economic impact estimated by multiplier-based methods needs to be offset by some corresponding lost economic impact that was crowded out, or else the net economic impact will be overstated.

It is extremely difficult to determine how much crowding out actually takes place when an area hosts the Olympic Games. However, one study found that gate arrivals at the Atlanta airport during the 1996 Summer Games were identical to gate arrivals in the same months in 1995 and 1997, implying that quite a few tourists to Atlanta were crowded out by the 1996 Games (Porter 1999). In late 2004, Athens tourism officials were estimating about a 10% drop in summer tourism in 2004 due to the Olympics. The Utah Skier Survey found that nearly 50% of non-residents would stay away from Utah in 2002 due to the expectation of more crowds and higher prices. The Beijing Tourism Bureau projected that the number of visitors to the city in August 2008 during the Games would not be greater than in August 2007. An additional problem for Beijing was that in order to abate the city's intense pollution during the summer months, the government ordered many of the city's factories closed leading up to and during the Games, and it imposed severely restrictive driving regulations.

Third, multiplier-based estimates overlook the displacement phenomenon. Some local residents may choose to leave town to avoid the congestion during the Games. The displaced people spend money outside the local area that they would have spent locally absent the Games. For instance, a survey in Barcelona indicated that fully one-sixth of the city's residents planned to travel outside the city during the 1996 Olympics.

Fourth, multiplier-based estimates depend critically on the selection of the multiplier. Economic theory does not provide exact guidance on the size of the multiplier to use in any particular application. The size of the multiplier used is at the discretion of the analyst. This creates an incentive for researchers to systematically choose large multipliers in order to generate large estimates of the economic impact of sporting events.

Despite all these problems, the majority of published estimates of the economic impact of the Olympic Games come from multiplier-based estimates. Multiplier-based estimates are widely used because, relative to the other approaches discussed above, this approach requires little data, little technical expertise, and very little in the way of computing power. Multiplier-based estimates are relatively cheap to produce and easy to manipulate.

Considering the size and prominence of the event, relatively little objective evidence on the economic impact of the Olympic Games exists. Much of the existing evidence has been developed by the host cities or regions, which have a vested interest in justifying the large expenditures on the games that were documented above. These 'promotional' studies, which have produced widely disparate estimated impacts of between $40 million and $16 billion for different Olympic Games between 1984 and 2006, suffer from the flaws discussed, and should be viewed sceptically. For a more detailed discussion of these studies, see Humphreys and Zimbalist (2008).

Estimates of the economic impact of the Olympic Games derived from academic research published in peer-reviewed journals tend to be more reliable. Only a few such studies, however, exist (Hotchkiss et al. 2003; Jasman and Maennig 2008; Feddersen and Maennig 2008; Lybbert and Hilmany 2000; Teigland 1999; Ritchie and Smith 1991).

The results of these studies present a consistent picture of the economic impact of hosting the Olympic Games on regions. Some jobs will be created as a result of hosting the Games. However, there appears to be no detectable effect on income,

suggesting that existing workers do not benefit from the Games. Moreover, the overall economic impact of hosting the Games depends on the overall labour market response to the new jobs created by the Games. When taking into account the overall labour market situation, the net impact of the Games on a region may not be positive. The negative impact on regional income found by the study that examined four North American regions is consistent with a negative overall labour market response to hosting the Games. Furthermore, the long-run impacts on tourism in the host region may be overstated, based on evidence from Lillehammer. Clearly, the weak results from academic research on the economic impact of hosting the Olympic Games call into question the reported economic impact from the promotional studies.

Some economists also have looked beyond income and employment measures for evidence that hosting the Olympic Games has an economic impact on the host economy. One area examined is stock markets (Berman et al. 2000; Veraros et al. 2004). The relationship between hosting the Olympic Games and stock markets is straightforward. To the extent that hosting the Olympic Games generates any benefits, including tangible economic benefits associated with increased tourism, or intangible benefits like national pride, sporting benefits, increased visibility etc., stock markets should be efficient mechanisms for valuing these benefits far into the future and discounting them back to the present. Positive benefits, if present, may be capitalized into stock prices at the time that the Games are awarded.

Research on the effect of hosting the Olympic Games on stock markets exploits the nature of the process through which the Games are awarded. Until the announcement of the winning city is made, there is considerable uncertainty about who will be awarded the games, and the contest is winner-take-all. The announcement about the winner of the Games takes place at a specific time (seven years prior to the Games) and represents a natural experiment in stock prices.

The existing evidence is mixed. The announcement that Sydney would host the 2000 Summer Olympic Games produced modest increases in stock returns in a limited number of industries: building materials, developers and contracts, and engineering. The announcement that Athens would host the 2004 Summer Olympic Games produced a short-term, significant increase in overall stock returns on the Athens Stock Exchange, but had no impact on the Milan Stock Exchange. Milan was one of the cities in the running for the 2004 Summer Games. Stock returns in construction related industries on the Athens Stock Exchange increased more than other sectors following the announcement, suggesting that much of the economic benefit accrues to this sector.

This evidence is limited to only two Olympic Games, and the increases in stock returns reported in the studies are modest, short-term, and primarily limited to the construction industry and related sectors of the economy. The empirical model used to analyze stock returns on the Athens Stock Exchange explains only 6% of the observed variation in returns. Overall, the evidence from this literature suggests that stock markets do not forecast large positive economic impacts flowing from the Olympic Games. While the idea that hosting the Olympic Games affects stock returns may appear important to the general public, a careful reading of this literature reveals that the underlying effects are small, transitory and limited to a few sectors of the economy. This evidence does not support net economic impact to the host city or region on the order of those publicized in promotional reports. Further, to the extent that hosting the Games may produce substantial fiscal deficits and growing public debt, the long-term effect on securities markets may well be negative.

Can the Olympic Games Be an Economic Success?

This review reveals relatively little evidence that hosting the Games produces significant economic benefits for the host city or region. If the economic gains are modest, or perhaps non-existent, what can host cities and regions do to maximize the potential gains from hosting the Olympic Games? A careful examination of past experiences suggests two important avenues for leveraging the

Olympic Games: host cities or regions need to make careful land use decisions and exploit the post-Olympic Games use of new and renovated facilities and infrastructure.

Land is an increasingly scarce resource both in the large urban areas that typically host the Summer Games and in the mountainous areas that host the Winter Games. Hosting the Olympic Games requires a significant amount of land for facilities, the Olympic Village, housing for the media and staff, accommodations for spectators, and parking. Unsuccessful Games leave behind legacies of seldom or never used structures taking up valuable land.

Successful Games, like the 1984 Los Angeles Summer Games, utilize existing facilities as much as possible, consuming as little scarce urban land as possible. The stadium used for the opening and closing ceremonies in the 1996 Atlanta Games was reconfigured to a baseball stadium immediately following the conclusion of the Games. The bullet train built for the Nagano Games greatly reduced the travel time between that city and Tokyo.

Tying up scarce land for seldom-used Olympic venues in both urban areas and alpine recreation areas cannot be an optimal use of this valuable resource. Olympic planners need to design facilities that will be useful for a long time after the Games are over, and are constructively integrated into the host city or region.

Clearly, the impact of the Olympic Games will vary according to the differing levels of development in the host city and country. Properly planned, hosting the Games can catalyze the construction of a modern transportation, communications and sport infrastructure. Such a potential benefit is bound to be greater for less developed areas. But even in such areas, hosting the Games will require a significant outlay of public funds to finance the infrastructural improvements. These improvements can also be made without hosting the Games. Thus it is relevant to ask whether the planning for the Olympics produces an optimal use of scarce public monies. It is also relevant to consider that in many circumstances the public policy process is so gridlocked that needed infrastructural investments may be delayed for years, if not decades, without the Olympic catalyst and that the Games do provide at least some capital to facilitate the completion of desirable projects.

Conversely, in more developed regions, where land is even more scarce during the initial bid planning (and destined to become scarcer still over the ten-year period of Olympic planning, bidding, selection and preparation) and labour and resource markets are tight, hosting the Games can occasion a gross misuse of land as well as provoke wage and resource price pressure leading to higher inflation.

Finally, it is important to recognize that hosting the Olympic Games may generate significant non-pecuniary benefits to the host city or region. The residents of the host city or region are likely to derive significant pride and sense of community from hosting the Games. Their homes are the focus of the world's attention for a brief but intense period. The planning and work required to host the Games takes considerable time and effort, and much of the hard work is done by volunteers. Pulling off such a huge endeavour is a source of considerable local and national pride. These factors are both important and valuable, even though researchers find it difficult to place a dollar value on them.

Some recent research has attempted to quantify the value of the non-pecuniary benefits generated by the Olympic Games (Atkinson et al. 2008). Economists have used the Contingent Valuation Method (CVM) to place a dollar value on such diverse intangible benefits as cleaning up oil spills in pristine wilderness areas and preserving green space in urban areas. The basic approach in CVM is to elicit people's willingness to pay for some intangible through hypothetical questions involving referendum voting or changes in taxes. A recent estimate of the total willingness to pay for the intangible benefits generated in the United Kingdom from hosting the 2012 Summer Games was in excess of d2 billion.

In the end, the economic and non-economic value to hosting the Olympic Games is a complex matter, likely to vary from one situation to another. Simple conclusions are impossible to draw. Prospective hosts of future Games would do well to steer clear of the inevitable Olympic

hype and to take a long, hard and sober look at the long-run development goals of their region.

This article is adapted from Chapter 6 in Zimbalist, A. 2010. *Circling the Bases: Essays on the Challenges and Prospects of the Sports Industry*. Temple University Press, Philadelphia.

See Also

► Sports, Economics of

Bibliography

Ahlfeldt, G., and W. Maennig. 2009. Impact of sports arenas on land values: Evidence from Berlin. *The Annals of Regional Science* 44(2): 205–227.

Atkinson, G., S. Mourato, and S. Szymanski. 2008. Are we willing to pay enough to 'back the bid?': Valuing the intangible benefits of hosting the summer Olympic Games. *Urban Studies* 45: 419–444.

Austen, I. 2010. A $1 billion hangover from an Olympic party. *New York Times*, 25 February.

Bartlett, D.L., and J.B. Steele. 2001. Snow job. *Sports Illustrated* 21(December): 79–98.

Berman, G., R. Brooks, and S. Davidson. 2000. The Sydney Olympic Games announcement and Australian stock market reaction. *Applied Economics Letters* 7: 781–784.

Burton, R. and O'Reilly, N. 2009. Consider intangibles when weighing Olympic host city benefits. *Sports Business Journal*, 7–13 September, p. 33.

Carlin, B. 2007. Olympic budget trebles to d9.3bn. *Daily Telegraph*, 15 March. *Daily Telegraph* 2008. 13 November.

Crompton, J. 1995. Economic impact analysis of sports facilities and events: Eleven sources of misapplication. *Journal of Sport Management* 9: 15.

Essex, S., and B. Chalkley. 2004. Mega-sporting events in urban and regional policy: A history of the Winter Olympics. *Planning Perspectives* 19(1): 205.

Humphreys, B., and A. Zimbalist. 2008. The financing and economic impact of the Olympic Games. In *The business of sports*, ed. B.R. Humphreys and D.R. Howard. Westport: Praeger.

International Olympic Committee. 2007. 2006 Olympic Marketing Fact File. http://www.olympic.org/uk/organisation/facts/revenue/index_uk.asp. Accessed 28 Oct 2007.

Porter, P. 1999. Mega-sports events as municipal investments: A critique of impact analysis. In *Sports economics: Current research*, ed. J. Fizel, E. Gustafson, and L. Hadley. Westport: Prager.

Preuss, H. 2003. *The economics of staging the Olympic Games*. Northampton: Edward Elgar Publishing.

Rephann, T., and A. Isserman. 1994. New highways as economic development tools: An evaluation using quasi-experimental matching methods. *Regional Science and Urban Economics* 24(6): 728.

Simon, B. 2006. Cost of Canadian 2010 Winter Olympics escalates. *Financial Times*, 6 February.

Sports Business Daily. 2008a. 14 November.

Sports Business Daily. 2008b. 14 November.

Sports Business Daily. 2009a. 27 October.

Sports Business Daily. 2009b. 14 January.

Sports Business Daily. 2009c. 30 September.

Sports Business Journal. 2009a. 7–13 September.

Sports Business Journal. 2009b. 10–18 October, p. 8.

Ungar, B.L. 2000. *Olympic Games: Federal government provides significant funding and support*. Collingwood: Diane Publishing Co..

Veraros, N., E. Kasimati, and D. Dawson. 2004. The 2004 Olympic Games announcement and its effect on the Athens and Milan stock exchanges. *Applied Economics Letters* 11: 749–753.

Economic Integration

Bela Balassa

In everyday parlance, integration is defined as bringing together of parts into a whole. In the economic literature, the term 'economic integration' does not have such a clear-cut meaning. At one extreme, the mere existence of trade relations between independent national economies is considered as a form of economic integration; at the other, it is taken to mean the complete unification of national economies.

Economic integration is defined here as process and as a state of affairs. Considered as a process, it encompasses measures designed to eliminate discrimination between economic units that belong to different national states; viewed as a state of affairs, it represents the absence of various forms of discrimination between national economies.

Economic integration may take several forms that represent various degrees of integration. In a free trade area, tariffs (and quantitative import restrictions) among participating countries are eliminated, but each country retains its own tariffs against non-members. Establishing a customs

union involves, apart from the suppression of intra-area trade barriers, equalizing tariffs on imports from non-member countries.

A common market goes beyond a customs union, inasmuch as it also entails the free movement of factors of production. In turn, an economic union combines the suppression of restrictions on commodity and factor movements with some degree of harmonization of national economic policies, so as to reduce discrimination owing to disparities in these policies. Finally, total economic integration means the unification of economic policies, culminating in the establishment of a supra-national authority whose decisions are binding for the member states.

History

The first important case of economic integration was the German Zollverein in the 19th century, which subsequently led to total economic integration through the unification of the German states with the establishment of the Deutsches Reich. In the 20th century, the creation of the Benelux customs (1948) and subsequently economic (1949) union, comprising Belgium, Luxemburg, and the Netherlands, represented the first step towards European economic integration. It was followed by the establishment of the European Coal and Steel Community (1953) and the European Economic Community or EEC (1958), both comprising Belgium, France, Italy, Luxemburg, the Netherlands, and West Germany.

Austria, Denmark, Norway, Portugal, Sweden, Switzerland, and the United Kingdom founded the European Free Trade Association or EFTA in 1960, with Finland participating first as an associate and later as a full member. In turn, Denmark and the United Kingdom left EFTA and, together with Ireland, entered the European Economic Community in 1968; Greece became a member of the EEC in 1978, and Portugal and Spain joined in 1986.

In Eastern Europe, the Council for Mutual Economic Assistance or CMEA was established in 1948, with the participation of the Soviet Union, Bulgaria, Czechoslovakia, Hungary, Poland, and Romania. Albania and East Germany joined shortly thereafter; subsequently, Cuba and Mongolia became full members while Albania ceased to participate in CMEA activities.

There have been a number of attempts at economic integration in developing countries. Some were to involve the establishment of a free trade area, such as the Latin American Free Trade Association (1960) comprising Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela; others were designed to become customs unions, such as the West African Customs Union (1959), including the Ivory Coast, Mali, Mauritania, Niger, Senegal, and Upper Volta. In 1960, the Central American Common Market was established, with Costa Rica, Guatemala, Honduras, Nicaragua, and El Salvador as members; in turn, the East African Common Market, comprising Kenya, Tanzania, and Uganda and subsequently transformed into the East African Economic Community (1967), was designed to become an economic union. None of these attempts has come to fruition, however, as barriers to intra-area trade have not been fully eliminated or have subsequently been restored.

Trade Creation and Trade Diversion

Viner's The Customs Union Issue (1950) was the first important contribution to the theory of economic integration. Viner investigated the impact of a customs union on trade flows and distinguished between the 'trade-creating' and the 'trade-diverting' effects of a union. In the first case, there is a shift from domestic to partner country sources of supply of a particular commodity; in the second case, the shift occurs from non-member country to partner country sources of supply.

Trade creation increases economic welfare, inasmuch as higher-cost domestic sources of supply are replaced by lower-cost imports from partner countries that were previously excluded by the tariff. In turn, trade diversion has a welfare cost since tariff discrimination against non-member countries, attendant on the establishment of the customs union, leads to the replacement of lower-cost sources of supply in

these countries by higher-cost partner country sources.

The net welfare effects of the customs union will depend on the amount of trade created and diverted as well as on differences in unit costs. In a partial equilibrium framework, under constant costs, there will be a welfare gain (loss) if the amount of trade created, multiplied by differences in unit costs between the home and the partner countries, exceeds (falls short of) the amount of trade diverted, multiplied by differences in unit costs between the partner and the non-member countries.

Meade (1955) further considered the effects of a customs union on intercommodity substitution, involving the replacement of domestic products by partner country products (trade creation) and the replacement of products of non-member countries by partner country products (trade diversion). As in the case of substitution among the sources of supply of a particular commodity (production effects), trade creation involves a welfare improvement, and trade diversion the deterioration of welfare, in the event of substitution among commodities (consumption effects).

The separation of production and consumption effects does not imply the absence of interaction between the two. Substitution among sources of supply will affect the pattern of consumption through changes in the prices paid by the consumer. Also, intercommodity substitution will lead to modifications in the pattern of production by changing the prices received by producers.

At the same time, as Lipsey and Lancaster (1956–57) first noted, production and consumption effects – and the theory of customs unions in general – should be considered as special cases of the theory of the second best. Assuming that the usual conditions for a Pareto optimum are fulfilled, free trade will lead to efficient resource allocation while pre-union, as well as the post-union, situations are sub-optimal because tariffs exist in both cases. In the abstract, then, one cannot make a judgement as to whether establishing a customs union will increase or reduce welfare. Nevertheless, a consideration of certain factors may provide a presumption as to the possible direction of the welfare effects of a union.

Factors Influencing the Welfare Effects of a Custom Union

Lipsey (1960) suggested that the welfare effects of a customs union will depend on the relative importance in home consumption of goods produced domestically and imported from non-member countries prior to the establishment of the union. Ceteris paribus, the larger the share of domestic goods and the smaller the share of goods imported from non-member countries, the greater is the likelihood of an improvement in welfare following the union's establishment. Such will be the case since substitution of partner country products for domestic products entails trade creation and their substitution for the products of non-member countries involves trade diversion.

These propositions are consistent with Tinbergen's (1957) conclusion that increases in the size of a customs union will augment the probability of favourable welfare effects; in the limiting case, the customs union includes the entire world, which is equivalent to free trade. Applying the argument that gains are obtained through the enlargement of a union because of increased possibilities for the reallocation of production, it also follows that the gains are positively correlated with increases in the market size of the participating countries (e.g. small countries will gain more from participation in a customs union than large countries).

Viner further considered the implications that differences in production structures among the member countries have for the welfare effects of a customs union. He suggested that the more competitive (the less complementary) is the production structure of the member countries, the greater is the chance that a customs union will increase welfare.

This proposition reflects the assumption that countries with similar production structures tend to replace domestic goods by competing imports from partner countries following the establishment of a customs union, while differences in the production structure within the union lead to substitution of partner country products for lower-cost products originating in non-member

countries (the latter conclusion does not hold if the union includes the low-cost producer).

The welfare effects of a customs union will also depend on transportation costs. Ceteris paribus, the lower are transport costs among the member countries, the greater will be the gains from their economic integration. Thus, the participation of neighbouring countries in a union, with greater possibilities for trade creation across their borders, will offer advantages over the participation of faraway countries that tends to promote trade diversion.

The height of tariffs will further affect the potential gains and losses derived from a customs union. High pre-union tariffs against the future member countries will increase the possibility of trade creation, and hence gains in welfare, following the establishment of the union while low tariffs against non-member countries will reduce the chances for trade diversion. But, these conclusions have little relevance under the application of the most-favoured-nation clause that entails providing equal tariff treatment to all countries before the customs union is established.

Customs Unions vs. Unilateral Tariff Reductions

In the Viner-Meade-Lipsey analysis, participation in a trade-creating customs union was considered as a means to reduce the distorting effects of the country's own tariffs. This argument was carried to its logical conclusion in contributions by Cooper and Massell (1965a) and Johnson (1965) who suggested that participation in a customs union is inferior to the unilateral elimination of tariffs, which leads to greater trade creation without giving rise to trade diversion.

The same authors claimed that the reasons for the establishment of customs unions lie in the gains participating countries may obtain in furthering non-economic objectives, and considered preference for industry as such an objective. They further assumed that this objective can be pursued at a lower cost in the framework of the larger market of a customs union than in the country's own domestic market.

As Johnson noted, the formation of a customs union in the pursuit of the stated objective presupposes that the member countries are at a comparative disadvantage in the production of industrial goods vis-à-vis the rest of the world. Cooper and Massell (1965b) identified such countries with developing countries, further suggesting that the economic planners of these countries are willing to accept some reduction in national income in order to assure increases in industrial production.

The question remains as to why there is a preference for industry. Johnson (1965) expressed the view that such preference may reflect nationalist aspirations and rivalry with other countries; the power of industrial firms and workers to increase their incomes; or the belief that industrial activity involves beneficial externalities. The last point, however, implies that there is no need to introduce non-economic considerations to obtain the Cooper–Mansell–Johnson result; the desirability of a customs union may be established in economic terms, provided that it permits obtaining externalities that cannot be achieved otherwise.

A further question is if unilateral tariff reductions will be superior to a customs union in the absence of a preference for industry or beneficial externalities. The Wonnacotts (1981) showed that this may not be the case if one admits the existence of tariffs in partner and in non-member countries prior to the formation of the customs union.

The elimination of tariffs by partner countries will provide benefits to the home country as it can now sell at a higher price in partner country markets. This gain will be larger the higher is the pre-union tariff in the partner countries and will further be affected by tariffs in the non-member countries. This is because, in selling in partner country markets free of duty, home country producers avoid paying the tariff in non-member countries.

Finally, Cooper and Massell (1965b) noted that a subsidy-union, with each participating country subsidizing its own industrial production, is superior to a customs union. This conclusion follows since the consumption cost of the tariff can be

avoided if the prices of industrial products in the union are maintained at the world market level through subsidies. However, production subsidization may be done by each country individually, with the attendant welfare benefits, without participating in a union.

Multi-country Analysis of a Customs Union

Traditionally, the welfare effects of a customs union were considered from the point of view of a single country. Yet, these effects may differ among member countries, depending on their production structure, location, the height of pre-union tariffs, and other characteristics. In fact, one member country may obtain a gain and another a loss, when any attempt to aggregate gains and losses encounters the well-known difficulties of international welfare comparisons.

The distribution of welfare gains and losses in a customs union will be further affected by changes in the terms of trade. The establishment of a union may give rise to price changes in trade between the member countries, even if the prices at which trade takes place with non-member countries remain unchanged (the case of the 'small' union).

In the more general case, prices in trade with non-member countries will also vary. Now, while trade diversion involves a welfare loss to the member countries of a customs union under unchanged terms of trade, this loss may be offset by a welfare gain due to improvements in the terms of trade attendant on trade diversion. Conversely, whereas under the assumption of unchanged terms of trade the welfare of non-member countries is unaffected by the establishment of a customs union, non-member countries will lose owing to the adverse impact of trade diversion on their terms of trade. This may be interpreted as the result of a shift in the union members' reciprocal demand curve for products originating in non-member countries.

Improvements in the terms of trade thus provide reasons for the establishment of a customs union even in the absence of non-economic objectives and beneficial externalities. Such improvements also favour a customs union over unilateral tariff reductions, which would lead to the deterioration of the terms of trade of the country concerned.

Other things being equal, the larger the union the greater will be its gain, and hence the loss to non-member countries, through terms of trade changes. This is because, ceteris paribus, the larger the union the higher will be the elasticity of its reciprocal demand for foreign products and the lower the elasticity of reciprocal demand on the part of non-member countries for the union's products.

The extent of terms-of-trade effects will further depend on the height of tariffs before and after the establishment of a customs union. As Vanek (1965) first showed, a customs union will not involve a loss to non-member countries, while benefiting its own members, if the union's external tariff level is sufficiently lower than the pre-union tariffs of the member countries.

Vanek's proposition was formulated in a three-country, two-commodity (3×2) model. It has subsequently been extended to a general case, under which compensatory payments to non-member countries were also introduced (Kemp and Wan 1976). At the same time, these propositions indicate a theoretical possibility rather than a likely outcome, since customs unions have shown little inclination to compensate non-member countries for losses attendant upon the union's establishment.

3×3 models represent an intermediate case between 3×2 and $m \times n$ models. They permit introducing a greater number of possible trade patterns, differential tariffs, complementarity and substitution in consumption, with a large number of marginal conditions in production and consumption, as well as intermediate products (Lloyd 1982). The 3×3 model is thus richer in content than the 3×2 model. Despite attempts made at introducing new terminology (Collier 1979), however, adding a third commodity does not appear to have materially affected the basic propositions of customs union theory. This conclusion may also find application to $m \times n$ models.

Free Trade Areas

In a free trade area, maintaining different tariffs among member countries on the products of non-members introduces the possibility of trade deflection. Furthermore, production and investment deflection may occur if one admits trade in intermediate products.

There will be trade deflection if imports enter the free trade area via the member country which applies the lowest tariff. Transportation costs apart, this is equivalent to adopting a tariff equal to the lowest tariff for each commodity in any of the member countries. Under the assumption of unchanged terms of trade, the deflection of trade will increase welfare in the member countries by limiting the extent of trade diversion. Removing this assumption, trade deflection will affect the distribution of welfare between member and non-member countries by reducing the terms of trade gain (loss) for the former (latter).

Production deflection will occur if the manufacture of products containing imported inputs shifts to countries which have lower tariffs on these inputs, because differences in tariffs outweigh differences in production costs. The deflection of production will have unfavourable effects on welfare, since the pattern of productive activity will not follow lines of comparative advantage but rather differences in duties.

The deflection of production may also affect the pattern of investment. Other things being equal, investors will establish factories in countries with lower tariffs on imported inputs. Again, adverse welfare effects will ensue because investments respond to tariff differences rather than to differences in production costs.

The deflection of trade, production, and investment represent unintended effects of free trade areas. To avoid such an eventuality member countries of free trade areas have imposed country of origin rules. These rules limit the freedom of intra-area trade to commodities that incorporate a certain proportion of domestic products or undergo a particular process of transformation in one of the member countries. The application of origin rules limits, but does not entirely eliminate, trade, production and investment deflection in a free trade area. Other things being equal, then, their self-interest would tend to encourage member countries to reduce their own tariffs.

Factor Movements

The deflection of investment may occur within a country or may involve international capital movements. In the first case, it affects the allocation of the country's own capital among industries; in the second case, it influences the international allocation of capital.

The last point leads to the case of common markets where, by definition, the full mobility of factors is assured. Meade (1953) first analysed the welfare effects of the movement of factors of production in an integrated area. He concluded that free factor movement will increase the gains obtained in a union by reducing the relative scarcities of the factors of production. This conclusion reflected the assumption that the conditions for factor price equalization through trade are not fulfilled.

If factors of production were not free to move between member and non-member countries, there will be no welfare loss due to factor movements among member countries to correspond to trade diversion in commodity trade. In the event of such factor movements, however, an analogous case to trade creation and trade diversion occurs if the movement of factors were subject to taxes prior to the establishment of a union and these taxes have been removed among union members. And, in any case, there will be indirect effects on welfare to the extent that factor movements substitute for trade. These effects may involve welfare losses to non-member countries as the newly-established productions substitute for imports from them.

Economies of Scale

Economic integration may lead to lower costs through increases in the volume of plant output. For various types of equipment, such as containers, pipelines, and compressors, cost is a

function of the surface area whereas capacity is related to volume; per unit costs decline with increases in output in the case of bulk transactions as well as for nonproportional activities such as design production planning, research, and the collection and channelling of information; inventory holdings do not need to increase proportionately with output; larger output warrants the application of technological methods that call for the use of specialized equipment or assembly-line production; and large-scale production may be necessary to ensure the optimum use of various kinds of indivisible equipment.

Corden (1972) showed that the traditional concepts of trade creation and trade diversion will be relevant in the case of economies of scale on the plant level but new concepts are added: the cost-reduction effect and the trade-suppression effect. The former refers to reductions in average unit costs as domestic output expands following the establishment of the union; the latter refers to the replacement of cheaper imports from non-member countries by domestic production under economies of scale. In Corden's view, a net benefit is likely to ensue as the cost-reduction effect tends to outweigh the trade-suppression effect.

Plant size and unit costs are not necessarily correlated in the case of multiproduct firms. In such instances, costs may be lowered by reducing product variety through specialization in an integrated area, which permits lengthening production runs for individual products.

The advantages of longer production runs derive from improvements in manufacturing efficiency along the 'learning curve' as cumulated output increases; the lowering of expenses involved in moving from one operation to another that involves the resetting of machines, the shifting of labour, and the reorganization of the work process; and the use of special-purpose machinery in the place of general purpose machinery.

Apart from product or horizontal specialization, there are possibilities for vertical specialization by subdividing the production process among individual establishments in an integrated area. As the sales of the final product increase, parts, components, and accessories may be manufactured in separate plants, each of which enjoys economies of scale, thereby resulting in cost reductions.

Competition and Technological Change

Economic integration will also create the conditions for more effective competition (Scitovsky 1958). By increasing the number of firms each producer considers as his competitors, the opening of national frontiers will contribute to the loosening of monopolistic and oligopolistic market structures in the individual countries. At the same time, there is no contradiction between gains from economies of scale and increased competition, since a wider market can sustain a larger number of efficient units (Balassa 1961).

Greater competition may have beneficial effects through improvements in manufacturing efficiency as well as through technological change. While the former has no place in traditional theory, which postulates the choice of the most efficient production methods among those available to the firm, it may assume considerable importance in countries whose markets have been sheltered from foreign competition.

The stick and the carrot of competition also provides inducement for technological progress in the member countries. In particular, increased competition may stimulate research activity aimed at developing new products and improving production methods. Finally, economic integration may contribute to the transmission of technological knowledge by increasing the familiarity of producers with new products and technological processes originating in the partner countries.

It has been suggested, however, that gains from competition, and from economies of scale, may be obtained through unilateral trade liberalization and that the gains are predicated on the response of economic agents to the stimulus provided by competition (Krauss 1972). While the validity of the second point depends on factors which are particular to each country, the first neglects the gains obtained through the increases in output associated with sales in the markets of partner countries.

Policy Harmonization

Policy differences among the member countries may influence trade flows and factor movements, thereby modifying the welfare effects of economic integration. Industrial policies, social policies, fiscal policies, monetary policies, and exchange rate policies are relevant in this context (Balassa 1961).

Industrial policies may involve granting credit preferences and/or tax benefits across the board or to particular activities. 'Horizontal' policies that are applied across-the-board do not create distortions, unless the conditions under which they are provided favour one activity over another. By contrast, 'vertical' measures are granted to particular activities and thereby introduce distortions, which may counteract the effects of the elimination of intra-area tariffs.

Intercountry differences in social policies will not give rise to distortions, provided that social benefits are financed from the contributions of employers and employees. Nor are these conclusions affected if factor mobility is introduced into the analysis as long as the employees regard the resulting social benefits as part of their compensation.

The situation is different if social benefits are financed from general tax revenue. This case is equivalent to a wage subsidy that favours labour-intensive activities. Correspondingly, differences in the mode of financing social security among the member countries will introduce distortions in resource allocation. This conclusion is strengthened if consideration is given to factor movements that respond to international differences in labour costs.

The elimination of vertical measures of industrial policy and the equalization of the conditions of financing social security will reduce distortions in resource allocation as well as differences in tax burdens among the member countries. Differences in the tax burden may remain, however, owing to national preferences as to the provision of collective goods. The effects of such differences on factor movements will depend on the spending of the tax proceeds. But, there may be 'supply-side' effects, with a lower tax burden providing incentives for work effort and risk taking.

A further question is if, for a given tax burden, intercountry differences in reliance on indirect taxes and income taxes will distort competition. Under the destination principle, indirect taxes are rebated on exports and imposed on imports without such adjustments occurring in regard to income taxes. Nevertheless, distortions in the conditions of competition will not ensue as flexibility in exchange rates will offset differences in rates of indirect taxes.

The application of the origin principle, with indirect taxes levied on production irrespective of the country of sale, in one country and that of the destination principle in another will similarly be offset through exchange rate flexibility. Such will not be the case, however, if cascade-type taxation applied in one country and value added taxation in another, with the former raising the tax burden on industries that go through several stages of fabrication, each of which is subject to tax. Eliminating this source of distortion would necessitate the adoption of value added taxation in all member countries of a union.

While exchange rate flexibility is necessary to offset intercountry differences in systems of taxation, it has been proposed that fixed exchange rates be established following the creation of a union. But such an action is predicated on the coordination – and eventual unification – of monetary and fiscal policies, since otherwise pressures are created for exchange rate changes. The fixity of exchange rate should thus be considered as the final outcome of policy coordination rather than an intermediate step in economic integration (Balassa 1975).

See Also

- ▶ Customs Unions
- ▶ List, Friedrich (1789–1846)

Bibliography

Balassa, B. 1961. *The theory of economic integration*. Homewood: Richard D. Irwin.

Balassa, B. 1975. Monetary integration in European Common Market. In *European economic integration*, ed. B. Balassa, 175–220. Amsterdam: North-Holland.

Collier, P. 1979. The welfare effects of a customs union: An anatomy. *Economic Journal* 83: 84–87.

Cooper, C.A., and B.F. Massell. 1965a. A new look at customs union theory. *Economic Journal* 75: 742–747.
Cooper, C.A., and B.F. Massell. 1965b. Towards a general theory of customs unions for developing countries. *Journal of Political Economy* 73: 461–476.
Corden, W.M. 1972. Economies of scale and customs union theory. *Journal of Political Economy* 80: 465–475.
Johnson, H.G. 1965. An economic theory of protectionism, tariff bargaining, and the formation of customs unions. *Journal of Political Economy* 73: 256–283.
Kemp, M.C., and H.Y. Wan Jr. 1976. An elementary proposition concerning the formation of customs unions. *Journal of International Economics* 6: 95–97.
Krauss, M.B. 1972. Recent developments in customs union theory: An interpretative survey. *Journal of Economic Literature* 10: 413–436.
Lipsey, R.G. 1960. The theory of customs unions: A general survey. *Economic Journal* 70: 496–513.
Lipsey, R.G., and K.J. Lancaster. 1956–7. The general theory of second best. *Review of Economic Studies* 24, 11–32.
Lloyd, P.J. 1982. The theory of customs unions. *Journal of International Economics* 12: 41–63.
Meade, J.E. 1953. *Problems of economic union*. London: Allen & Unwin.
Meade, J.E. 1955. *The theory of customs union*. Amsterdam: North-Holland.
Scitovsky, T. 1958. *Economic theory and Western European integration*. London: Allen & Unwin.
Tinbergen, J. 1957. Customs unions: Influence of their size on their effect. *Zeitschrift der gesamten Staatswissenschaft* 113: 404–414.
Vanek, J. 1965. *General equilibrium of international discrimination. The case of customs unions*. Cambridge: Harvard University Press.
Viner, J. 1950. *The customs union issue*. New York: Carnegie Endowment for International Peace.
Wonnacott, P., and R. Wonnacott. 1981. Is unilateral tariff reduction preferable to a customs union? The curious case of the missing foreign tariffs. *American Economic Review* 71: 704–714.

Economic Interpretation of History

Ernest Gellner

Marxism does not possess a monopoly of the economic interpretation of history. Other theories of this kind can be formulated – for instance that which can be found in the very distinguished work of Karl Polanyi, dividing the history of mankind into three stages, each defined by a different type of economy. If Polanyi is right in suggesting that reciprocity, redistribution and the market each defined a different kind of society, this is, in a way, tantamount to saying that the economy is primary, and thus his work constitutes a species of the economic interpretation of history. Nevertheless, despite the importance of Polanyi's work and the possibility of other rival economic interpretations, Marxism remains the most influential, the most important, and perhaps the best elaborated of all theories, and we shall concentrate on it.

One often approaches a theory by seeing what it denies and what it repudiates. This approach is quite frequently adopted in the case of Marxism, where it is both fitting and misleading. We shall begin by adopting this approach, and turn to its dangers subsequently.

Marxism began as the reaction to the romantic idealism of Hegel, in the ambience of whose thought the young Karl Marx reached maturity. This no doubt is the best advertised fact about the origin of Marxism. The central point about Hegelianism was that it was acutely concerned with history and social change, placing these at the centre of philosophical attention (instead of treating them as mere distractions from the contemplation of timeless objects, which had been a more frequent philosophical attitude); and secondly, it taught that history was basically determined by intellectual, spiritual, conceptual or religious forces. As Marx and Engels put it in The German Ideology, 'The Young Hegelians are in agreement with the Old Hegelians in their belief in the rule of religion, of concepts, of an abstract general principle in the existing world' (Marx and Engels 1845–1846, p. 5).

Now the question is – why did Hegel and followers believe this? If it is interpreted in a concrete sense, as a doctrine claiming that the ideas of men determined their other activities, it does not have a great deal of plausibility, especially when put forward as an unrestricted generalization. If it is formulated – as it was by Hegel – as the view that some kind of abstract principle or entity dominates history, the question may well be asked: what evidence do we have for the very existence of this mysterious poltergeist

allegedly manipulating historical events? Given the fact that the doctrine is either implausible or obscure, or indeed both, why were intelligent men so strongly drawn to it?

The answer to this may be complex, but the main elements in it can perhaps be formulated simply and briefly. Hegelianism enters the scene when the notion of what we now call culture enters public debate. The point is this: men are not machines. When they act, they do not simply respond to some kind of push. When they do something, they generally have an idea, a concept, of the action which they are performing. The idea or conception in turn is part of a whole system. A man who goes through the ceremony of marriage has an idea of what the institution means in the society of which he is part, and his understanding of the institution is an integral part of his action. A man who commits an act of violence as part of a family feud has an idea of what family and honour mean, and is committed to those ideas. And each of these ideas is not something which the individual had excogitated for himself. He took it over from a corpus of ideas which differ from community to community, and which change over time, and which are now known as culture.

Put in this way, the 'conceptual' determination of human conduct no longer seems fanciful, but on the contrary is liable to seem obvious and trite. In various terminologies ('hermeneutics', 'structuralism', and others) it is rather fashionable nowadays. The idea that conduct is concept-saturated and that concepts come not singly but as systems, and are carried not by individuals but by on-going historic communities, has great plausibility and force. Admittedly, those who propose it, in Hegel's day and in ours, do not always define their position with precision. They do not always make clear whether they are merely saying that culture in this sense is important (which is hardly disputable), or claiming that it is the prime determinant of other things and the ultimate source of change, which is a much stronger and much more contentious claim. Nonetheless, the idea that culture is important and pervasive is very plausible and suggestive, and Hegelianism can be credited with being one of the philosophies which, in its own peculiar language, had introduced this idea. It is important to add that Hegelianism often speaks of 'Spirit' in the singular; our suggestion is that this can be interpreted as culture, as the spirit of the age. This made it easy for Hegelianism to operate as a kind of surrogate Christianity: those no longer able to believe in a personal god could tell themselves that this had been a parable on a kind of guiding historical spirit. For those who wanted to use it in that way, Hegelianism was the continuation of religion by other means.

But Hegelianism is not exhausted by its sense of culture, expressed in somewhat strange language. It is also pervaded by another idea, fused with the first one, and one which it shares with many thinkers of its period: a sense of historical plan. The turn of the 18th and 19th centuries was a time when men became imbued with the sense of cumulative historical change, pointing in an upward direction – in other words, the idea of Progress.

The basic fact about Marxism is that it retains this second idea, the 'plan' of history, but aims at inverting the first idea, the romantic idealism, the attribution of agency to culture. As the two founders of Marxism put it themselves in The German Ideology (pp. 14–15),

> In direct contrast to German philosophy which descends from heaven to earth, here we ascend from earth to heaven … We set out from real active men, and on the basis of their real life-process we demonstrate the development of the ideological reflexes and echoes of this life-process … Morality, regligion, metaphysics, all the rest of ideology and their corresponding forms of consciousness, thus no longer retain the semblance of independence. They have no history, no development; but men, developing their material production and their material intercourse, alter, along with their real existence, their thinking and the products of their thinking. Life is not determined by consciousness, but consciousness by life.

Later on in the same work, the two founders of Marxism specify the recipe which, according to them, was followed by those who produced the idealistic mystification. First of all, ideas were separated from empirical context and the interests

of the rulers who put them forward. Secondly, a set of logical connections was found linking successive ruling ideas, and their logic is then meant to explain the pattern of history. (This links the concept-saturation of history to the notion of historic design. Historic pattern is the reflection of the internal logical connection of successive ideas.) Thirdly, to diminish the mystical appearance of all this, the free-floating, self-transforming concept was once again credited to a person or group of persons.

If this kind of theory is false, what then is true? In the same work a little later, the authors tell us:

> This sum of productive forces, forms of capital and social forms of intercourse, which every individual and generation finds in existence as something given, is the real basis of . . . the . . . 'essence of man'.. . . These conditions of life, which different generations find in existence, decide also whether or not the periodically recurring revolutionary convulsion will be strong enough to overthrow the basis of all existing forms. And if these material elements of a complete revolution are not present. . . then, as far as practical developments are concerned, it is absolutely immaterial whether the 'idea' of this revolution has been expressed a hundred times already . . . (p. 30).

The passage seems unambiguous: what is retained is the idea of a plan, and also the idea of primarily internal, endogenous propulsion. What has changed is the identification of the propulsion, of the driving force of the transformation. Change continues to be the law of all things, and it is governed by a plan, it is not random; but the mechanism which controls it is now identified in a new manner.

From then on, the criticisms of the position can really be divided into two major species: some challenge the identification of the ruling mechanism, and others the idea of historic plan. As the most dramatic presentation of Marxist development, Robert Tucker's Philosophy and Myth in Karl Marx (1961, p. 123) puts it:

> Marx founded Marxism in an outburst of Hegelizing. He considered himself to be engaged in . . . [an]. . . act of translation of the already discovered truth . . . from the language of idealism into that of materialism. . .. Hegelianism itself was latently or esoterically an economic interpretation of history. It treated history as 'a history of production' . . . in which spirit externalizes itself in thought-objects. But this was simply a mystified presentation of man externalizing himself in material objects.

This highlights both the origin and the validity or otherwise of the economic interpretation of history. Some obvious but important points can be made at this stage. The Hegel/Marx confrontation owes much of its drama and appeal to the extreme and unqualified manner in which the opposition is presented. This unqualified, unrestricted interpretation can certainly be found in the basic texts of Marxism. Whether it is the 'correct' interpretation is an inherently undecidable question: it simply depends on which texts one treats as final – those which affirm the position without restriction and without qualification, or those which contain modifications, qualifications and restrictions.

The same dilemma no doubt arises on the Hegelian side, where it is further accompanied by the question as to whether the motive force, the spirit of history, is to be seen as some kind of abstract principle (in which case the idea seems absurd to most of us), or whether this is merely to be treated as a way of referring to what we now term culture (in which case it is interesting and contentious).

One must point out that these two positions, the Hegelian and the Marxist, are contraries, but not contradictories. They cannot both be true, but they can perfectly well both be false. A world is easily conceivable where neither of them is true: a world in which social changes sometimes occur as a consequence of changes in economic activities, and sometimes as a consequence of strains and stresses in the culture. Not only is such a world conceivable, but it does really rather look as if that is the kind of world we do actually live in. (Part of the appeal of Marxism in its early days always hinged on presenting Hegel-type idealism and Marxism as two contradictories, and 'demonstrating' the validity of Marxism as a simple corollary of the manifest absurdity of strong versions of Hegelianism.) In this connection, it is worth noticing that by far the most influential (and not unsympathetic) sociological critic of Marx is Max Weber, who upholds precisely this kind of

position. Strangely enough, despite explicit and categorical denials on his own part, he is often misrepresented as offering a return to some kind of idealism (without perhaps the mystical idea of the agency of abstract concepts which was present in Hegel). For instance, Michio Morishima, in Why has Japan 'Succeeded'? (1982, p. 1), observes: 'Whereas Karl Marx contended that ideology and ethics were no more than reflections . . . Max Weber . . . made the case for the existence of quite the reverse relationship.' Weber was sensitive to both kinds of constraint; he merely insisted that on occasion, a 'cultural' or 'religious' element might make a crucial difference.

Connected with this, there is another important theoretical difference to be found in Weber and many contemporary sociologists. The idea of the inherent historical plan, which had united Hegel and Marx, is abandoned. If the crucial moving power of history comes from one source only, though this does not strictly speaking entail that there should be a plan, an unfolding of design, it nevertheless does make it at least very plausible. If that crucial moving power had been consciousness, and its aim the arrival at self-consciousness, then it was natural to conclude that with the passage of time, there would indeed be more and more of such consciousness. So the historical plan could be seen as the manifestation of the striving of the Absolute Spirit or humanity, towards ever greater awareness. Alternatively, if the motive force was the growth of the forces of production, then, once again, it was not unreasonable to suppose that history might be a series of organizational adjustments to expanding productive powers, culminating in a full adjustment to the final great flowering of our productive capacity. (Something like that is the essence of the Marxist vision of history.)

If on the other hand the motive forces and the triggers come from a number of sources, which moreover are inherently diverse, there is no clear reason why history should have a pattern in the sense of coming ever closer to satisfying some single criterion (consciousness, productivity, congruence between productivity and social ethos, or whatever). So in the Weberian and more modern vision, the dramatic and unique developments of the modern industrial world are no longer seen as the inevitable fulfilment and culmination of a potential that had always been there, but rather as a development which only occurred because a certain set of factors happened to operate at a given time simultaneously, and which would otherwise not have occurred, and which was in no way bound to occur. Contingency replaces fatality.

So much for the central problem connected with the economic interpretation of history. The question concerning the relative importance of conceptual (cultural) and productive factors is the best known, most conspicuous and best advertised issue in this problem area. But in fact, it is very far from obvious that it is really the most important issue, the most critical testing ground for the economic theory of history. There is another problem, less immediately obvious, less well known, but probably of greater importance, theoretically and practically. That is the relative importance of productive and coercive activities.

The normal associations which are likely to be evoked by the phrase 'historical materialism' do indeed imply the downgrading of purely conceptual, intellectual and cultural elements as explanatory factors in history. But it does not naturally suggest the downgrading of force, violence, coercion. On the contrary, for most people the idea of coercion by threat or violence, or death and pain, seems just as 'realistic', just as 'materialistic' as the imperatives imposed by material need for sustenance and shelter. Normally one assumes that the difference between coercion by violence or the threat of violence, and coercion by fear of destitution, is simply that the former is more immediate and works more quickly. One might even argue that all coercion is ultimately coercion by violence: a man or a group in society which coerces other members by controlling the food supply, for instance, can only do it if they control and defend the store of food or some other vital necessity by force, even if that force is kept in reserve. Economic constraint, it could be argued (as Marxists themselves argue in other contexts), only operates because a certain set of rules is enforced by the state, which may well remain in the background. But economic constraint is in this way parasitic on

the ultimate presence of enforcement, based on the monopoly of control of the tools of violence.

The logic of this argument may seem persuasive, but it is contradicted by a very central tenet of the Marxist variant of the economic theory of history. Violence, according to the theory, is not fundamental or primary, it does not initiate fundamental social change, nor is it a fundamental basis of any social order. This is the central contention of Marxism, and at this point, real Marxism diverges from what might be called the vulgar image possessed of it by non-specialists. Marxism stresses economic factors, and downgrades not merely the importance of conceptual, 'superstructural' ones, but equally, and very significantly, the role of coercive factors.

A place where this is vigorously expressed is Engels's 'Anti-Dühring' (1878):

> . . . historically, private property by no means makes its appearance as the result of robbery or violence. . . . Everywhere where private property developed, this took place as the result of altered relations of production and exchange, in the interests of increased production and in furtherance of intercourse – that is to say, as a result of economic causes. Force plays no part in this at all. Indeed, it is clear that the institution of private property must be already in existence before the robber can appropriate another person's property. . . Nor can we use either force or property founded on force to explain the 'enslavement of man for menial labour' in its most modern form – wage labour. . .. The whole process is explained by purely economic causes; robbery, force, and the state of political interference of any kind are unnecessary at any point whatever (Burns 1935, pp. 267–9).

Engels goes on to argue the same specifically in connection with the institution of slavery:

> Thus force, instead of controlling the economic order, was on the contrary pressed into the service of the economic order. Slavery was invented. It soon became the predominant form of production among all peoples who were developing beyond the primitive community, but in the end was also one of the chief causes of the decay of that system (ibid., p. 274).
>
> Engels a little earlier in the same work was on slightly more favourable ground when he discussed the replacement of the nobility by the bourgeoisie as the most powerful estate in the land. If physical force were crucial, how should the peaceful merchants and producers have prevailed over the professional warriors? As Engels puts it: 'During the whole of this struggle, political forces were on the side of the nobility. . .' (ibid., p. 270).

One can of course think of explanations for this paradox: the nobility might have slaughtered each other, or there might be an alliance between the monarchy and the middle class (Engels himself mentioned this possibility, but does not think it constitutes a real explanation) and so forth. In any case, valid or not, this particular victory of producers over warriors would seem to constitute a prima facie example of the non-dominance of force in history. The difficulty for the theory arises when the point is generalized to cover all social orders and all major transitions, which is precisely what Marxism does.

Engels tries to argue this point in connection with a social formation which one might normally consider to be the very paradigm of the domination by force: 'oriental despotism'. (In fact, it is for this very reason that some later Marxists have maintained that this social formation is incompatible with Marxist theory, and hence may not exist.) Engels does it, interestingly enough, by means of a kind of functionalist theory of society and government: the essential function, the essential role and duty, of despotic governments in hydraulic societies is to keep production going by looking after the irrigation system. As he puts it:

> However great the number of despotic governments which rose and fell in India and Persia, each was fully aware that its first duty was the general maintenance of irrigation throughout the valleys, without which no agriculture was possible (Burns 1935, p. 273).

It is a curious argument. He cannot seriously maintain that these oriental despots were always motivated by a sense of duty towards the people they governed. What he must mean is something like this: unless they did their 'duty', the society in question could not survive, and they themselves, as its political parasites, would not survive either. So the real foundation of 'oriental despotism' was not the force of the despot, but the functional imperatives of despotically imposed irrigation systems. Economic need, as in the case of slavery, makes use of violence for its own ends, but violence itself initiates or maintains nothing. This

interpretation is related to what Engels says a little further on. Those who use force can either aid economic development or accelerate it, or go against it, which they do rarely (though he admits that it occasionally occurs), and then they themselves usually go under: 'Where... the internal public force of the country stands in opposition to economic development... the contest has always ended with the downfall of the political power' (Burns 1935, p. 277).

We have seen that Engels's materialism is curiously functional, indeed teleological: the economic potential of a society or of its productive base somehow seeks out available force, and enlists it on its own behalf. Coercion is and ought to be the slave of production, he might well have said. This teleological element is found again in what is perhaps the most famous and most concise formulation of Marxist theory, namely certain passages in Marx's preface to A Contribution to 'The Critique of Political Economy' (1859):

> A social system never perishes before all the productive forces have developed for which it is wide enough; and new, higher productive relationships never come into being before the material conditions for their existence have been brought to maturity within the womb of the old society itself. Therefore, mankind always sets itself only such problems as it can solve; for when we look closer we will always find that the problem itself only arises when the material conditions for its solution are already present, or at least in the process of coming into being. In broad outline, the Asiatic, the ancient, the feudal, and the modern bourgeois mode of production can be indicated as progressive epochs in the economic system of society (Burns 1935, p. 372).

The claim that a new order does not come into being before the conditions for it are available, is virtually a tautology: nothing comes into being unless the conditions for it exist. That is what 'conditions' mean. But the idea that a social system never perishes before it has used up all its potential is both strangely teleological and disputable. Why should it not be replaced even before it plays itself out to the full? Why should not some of its potential be wasted?

It is obvious from this passage that the purposive, upward surge of successive modes of production cannot be hindered by force, nor even aided by it. Engels, in 'Anti-Dühring', sneers at rulers such as Friedrich Wilhelm IV, or the then Tsar of Russia, who despite the power and size of their armies are unable to defy the economic logic of the situation. Engels also treats ironically Herr Dühring's fear of force as the 'absolute evil', the belief that the 'first act of force is the original sin', and so forth. In his view, on the contrary, force simply does not have the capacity to initiate evil. It does however have another 'role in history, a revolutionary role'; this role, in Marxist words, is midwifery:

> ... it is the midwife of every old society which is pregnant with the new,... the instrument by the aid of which social movement forces its way through and shatters the dead, fossilized, political forms ... (Burns 1935, p. 278).

The midwifery simile is excellent and conveys the basic idea extremely well. A midwife cannot create babies, she can only aid and slightly speed up their birth, and once the infant is born the midwife cannot do much harm either. The most one can say for her capacity is that she may be necessary for a successful birth. Engels seems to have no fear that this sinister midwife might linger after the birth and refuse to go away. He makes this plain by his comment on the possibility of a 'violent collision' in Germany which 'would at least have the advantage of wiping out the servility which has permeated the national consciousness as a result of the humiliation of the Thirty Years War'.

There is perhaps an element of truth in the theory that coercion is and ought to be the slave of production. The element of truth is this: in pre-agrarian hunting and gathering societies, surrounded by a relative abundance of sustenance but lacking means of storing it, there is no persistent, social, economic motive for coercion, no sustained employment for a slave. By contrast, once wealth is systematically produced and stored, coercion and violence or the threat thereof acquire an inescapable function and become endemic. The surplus needs to be guarded, its socially 'legitimate' distribution enforced. There is some evidence to support the view that hunting and gathering societies were more peaceful than the agrarian societies which succeeded them.

One may put it like this: in societies devoid of a stored surplus, no surplus needs to be guarded and the principles governing its distribution do not need to be enforced. By contrast, societies endowed with a surplus face the problem of protecting it against internal and external aggression, and enforcing the principles of its distribution. Hence they are doomed to the deployment, overt or indirect, of violence of the threat thereof. But all of this, true though it is, does not mean that surplusless societies are necessarily free of violence: it only means that they are not positively obliged to experience it. Still less does it mean that within the class of societies endowed with a surplus, violence on its own may not occasionally or frequently engender changes, or inhibit them. The argument does not preclude coercion either from initiating social change, or from thwarting change which would otherwise have occurred. The founding fathers of Marxism directed their invective at those who raised this possibility, but they never succeeded in establishing that this possibility is not genuine. All historic evidence would seem to suggest that this possibility does indeed often correspond to reality.

Why is the totally unsubstantiated and indeed incorrect doctrine of the social unimportance of violence so central to Marxism?

The essence of Marxism lies in the retention of the notion of an historical plan, but a re-specification of its driving force. But the idea of a purposive historical plan is not upheld merely out of an intellectual desire for an elegant conceptual unification of historical events. There is also a deeper motive. Marxism is a salvation religion, guaranteeing not indeed individual salvation, but the collective salvation of all mankind. Ironically, its conception of the blessed condition is profoundly bourgeois. Indeed, it constitutes the ultimate apotheosis of the bourgeois vision of life. The bourgeois preference for peaceful production over violent predation is elevated into the universal principle of historical change. The wish is father of the faith. The work ethic is transformed into the essence, the very species-definition of man. Work is our fulfilment, but work patterns are also the crucial determinants of historical change. Spontaneous, unconstrained work, creativity, is our purpose and our destiny. Work patterns also determine the course of history and engender patterns of coercion, and not vice versa. Domination and the mastery of techniques of the violence is neither a valid ideal, nor ever decisive in history. All this is no doubt gratifying to those imbued with the producer ethic and hostile to the ethic of domination and violence: but is it true?

Note that, were it true, Marxism is free to commend spontaneously cooperative production, devoid of ownership and without any agency of enforcement, as against production by competition, with centrally enforced ground rules. It is free to do it, without needing to consider the argument that only competition keeps away centralized coercion, and that the attempt to bring about propertyless and total cooperation only engenders a new form of centralized tyranny. If tyranny only emerges as a protector of basically pathological forms or organization of work, then a sound work-pattern will on its own free us for ever from the need for either authority or checks on authority. Man is held to be alienated from his true essence as long as he works for extraneous ends: he finds his true being only when he indulges in work for the sake of creativity, and choses his own form of creativity. This is of course precisely the way in which the middle class likes to see its own life. It takes pride in productive activity, and chooses its own form of creativity, and it understands what it does. Work is not an unintelligible extraneous imposition for it, but the deepest fulfilment.

On the Marxist economic interpretation of history, mankind as a whole is being propelled towards this very goal, this bourgeois-style fulfilment in work without coercion. But the guarantee that this fulfilment will be reached is only possible if the driving force of history is such as to ensure this happy outcome. If a whole multitude of factors, economic, cultural, coercive, could all interact unpredictably, there could hardly be any historic plan. But if on the other hand only one factor is fundamental, and that factor is something which has a kind of vectorial quality, something which increases over time and inevitably points in one direction only (namely the augmentation of the productive force of man), then the necessary historical plan does after all have a firm,

unprecarious base. This is what the theory requires, and this is what is indeed asserted.

The general problem of the requirement, ultimately, of a single-factor theory, with its well-directed and persistent factor, is of course related to the problems which arise from the plan that Marxists discern in history. According to the above quotation from Marx, subsequent to primitive communism, four class-endowed stages arise, namely the Asiatic, the ancient, the feudal, and the modern bourgeois, which is said to be the last 'antagonistic' stage (peaceful fulfilment follows thereafter). Marxism has notoriously had trouble with the 'Asiatic' stage because, notwithstanding what Engels claimed, it does seem to exemplify and highlight the autonomy of coercion in history, and the suspension of progress by a stagnant, self-maintaining social system.

But leaving that aside, in order to be loyal to its basic underlying intuition of a guaranteed progression and a final happy outcome, Marxism is not committed to any particular number or even any particular sequence of stages. The factual difficulties which Marxist historiography has had in finding all the stages and all the historical sequences, and in the right order, are not by themselves necessarily disastrous. A rigid unilinealism is not absolutely essential to the system. What it does require (apart from the exclusiveness, in the last analysis, of that single driving force) is the denial of the possibility of stagnation, whether in the form of absolute stagnation and immobility, or in the form of circular, repetitive developments. If this possibility is to be excluded, a number of things need to be true: all exploitative social forms must be inherently unstable; the number of such forms must be finite; and circular social developments must not be possible. If all this is so, then the alienation of man from his true essence – free fulfilment in unconstrained work – must eventually be attained. But if the system can get stuck, or move in circles, the promise of salvation goes by the board. This would be so even if the system came to be stuck for purely economic reasons. It would be doubly disastrous for it if other factors, such as coercion, were capable of freezing it. The denial of any autonomous role for violence in history is the most important, and most contentious, element in the Marxian economic theory of history.

So what the Marxist economic interpretation of history really requires is that no non-economic factor can ever freeze the development of society, that the development of society itself be pushed forward by the continuous (even if on occasion slow) growth of productive forces, that the social forms accompanying various stages of the development of productive forces should be finite in number, and that the last one be wholly compatible with the fullest possible development of productive forces and of human potentialities.

The profound irony is that a social system marked by the prominence and pervasiveness of centralized coercion, should be justified and brought about by a system of ideas which denies autonomous historical agency both to coercion and to ideas. The independent effectiveness both of coercion and of ideas can best be shown by considering a society built on a theory, and one which denies the effectiveness of either.

Bibliography

Burns, E. 1935. *A handbook of Marxism*. London: Victor Gollancz.

Engels, F. 1878. Anti-Dühring. In Burns (1935).

Marx, K. 1859. A contribution to 'the critique of political economy'. In Burns (1935).

Marx, K., and F. Engels. 1845–1846. *The German ideology*. London: Lawrence & Wishart, 1940.

Morishima, M. 1982. *Why has Japan 'succeeded'? Western technology and the Japanese ethos*. Cambridge: Cambridge University Press.

Tucker, R.C. 1961. *Philosophy and myth in Karl Marx*. Cambridge: Cambridge University Press.

Economic Laws

Stefano Zamagni

Keywords

Ceteris paribus; Economic laws; Law of diminishing returns; Law of variable proportions; Marginal revolution; Natural law; Neo-

Austrian economics; Positive economics; Substitution; Tendency laws

JEL Classifications
B0

The social sciences, and economics in particular, separated from moral and political philosophy in the second half of the 18th century when the results of the myriad of intentional actions of people were perceived to produce regularities resembling the laws of a system. Both Physiocratic thought and Smith's *Wealth of Nations* reflect this extraordinary discovery: scientific laws thought to be found only in nature could also be found in society. This extension poses several problems. A serious one refers to the tension of combining individuals' freedom of action with the scientists' desire to discover the systematic aspects of the unintended and quite often unpredictable consequences of human action, that is, the desire to arrive at laws characterized by a certain degree of generality and permanence.

In the history of economic thought this fundamental tension has been solved in different ways. In the 18th century, the mechanistic ideal of the natural sciences, combined with the natural law idea of a harmonious order of nature, determined the way social phenomena were treated. There was a desire to discover the 'natural laws' of economic life and to formulate the natural precepts which rule human conduct.

The classical economists upheld the notion that natural laws are embedded in the economic process as beneficial laws, along with the belief in the existence of rules of nature capable of being discovered. Thus the belief that things could follow the beneficial 'natural course' only in a rationally organized society which it was a duty to create according to the precepts of nature. The economic system is the mechanism by which the individual is driven to fostering the prosperity of society while pursuing his private interest. Hence the automatic operation of the economic system may be combined with freedom of individual action. This is the core of the doctrine of economic harmony. Besides being causal laws of a mechanical type, the laws of nature are providentially imposed norms of conduct. In such a setting it would have been pointless to separate means and ends, since the implementation of natural laws is both an end and a means, and even more pointless to think of a tension between 'explaining' and 'understanding' economic behaviour. Causal and teleological, positive and normative, theoretical and practical started being seen as separate categories only when the economic discourse freed itself from the philosophy of natural law and all its implications.

Post-classical economics set out to be a science of the laws regulating the economic order and of the conditions allowing these laws to operate. It became the basis of a theory that, in Jevons's own terms, proposed to construct a 'social physics'. The view of a social world ordered according to transcendent ends was abandoned in favour of an ideal of objective knowledge of economic phenomena gained through a 'positive' study of the laws that regulate market activities. In so doing, neoclassical 'positive' economics solves the aforementioned tension by extrapolating the theoretical model of natural sciences to economics: economics is to produce the laws of motion similar to those of physics, chemistry, astronomy.

But what is a scientific law and which role do laws play within the logical positivist's perspective adopted by neoclassical economics? Laws provide the foundation of a deductive scientific method of inquiry. According to the deductive–nomological conception of explanation, due to C. Hempel, laws are universal statements not requiring reference to any one particular object or spatio–temporal location. To be valid, laws are constrained neither to finite populations nor to particular times and places; they are, in effect, expressions of natural stationarities. This interpretation of the notion of law provides the so-called covering-law model of explanation with an unquestionably firm inferential foundation. Deductive logic is employed to ensure the truth status of propositions and, since the deductions are (by hypothesis) predicated on true universal statements (laws), the empirical validity of these statements may be ascertained. However, what sort of constraints on economic discourse are

imposed by this positivistic structure? On the one hand this structure constitutes its object; on the other hand it generates specific economic questions together with their method of solution. Following the model of natural sciences and its success in controlling a natural world made up of objects and unvarying relations among them expressed in the form of laws, the neoclassical approach arrives at a study of regularities conceived of as specifying the nature of its objects.

To capture the different interpretations of the notion of law by classical and neoclassical economists let us refer to one of the most famous of economic laws: the law of diminishing returns, also known as the law of variable proportions. Studying agricultural production, Ricardo had noted that different quantities of labour, assisted by certain quantities of other inputs (farm tools, fertilizers, and so on), could be employed on a given piece of land, that is, it was possible to vary the proportions in which land and complex labour (labour assisted by other inputs) are employed. He accordingly arrived at the law which states that production increases resulting from equal increments in the employment of complex labour, while the quantity of land farmed remains constant, will initially be increasing and then decreasing. (To be sure, the first statement of the law is due to the Physiocratic economist Turgot.)

Three points deserve attention. First, Ricardo and classical authors in general offer no formal demonstration of this law. To them, it is basically an empirical law, on which no functional association between output and variable inputs can be built. Second, the classics' use of the law refers to their theories of distribution and development: as the supply of land in the whole system is fixed, sooner or later a point will be reached at which economic growth will come to a halt, notwithstanding any countervailing effects due to technical progress. Finally, the law presupposes a comparative statics framework: the pattern of the marginal products of complex labour refers to different observable equilibrium positions and not to hypothetical or virtual variations.

With the advent of the marginalist revolution, two subtle changes in the interpretation of the law took place. (*a*) The *de facto* elimination of the distinction between the extensive case (the case of the simultaneous cultivation of pieces of land of different fertility) and the intensive case (the application of successive doses of capital and labour to the same piece of land) with an overevaluation of the latter. Classical economists, being interested in the explanation of rent, concentrated on the extensive case; they took also the intensive case into consideration but with many qualifications. Indeed, whereas the various levels of productivity of different qualities of land is a circumstance which may be directly observed in a given situation, the marginal productivity of a given input is related to a virtual increment in output and therefore to a virtual change in the situation. (*b*) The change in the method of analysis – it was preferred to reason in terms of hypothetical rather than observable changes – brought about by the shift of interest towards the intensive margin, supported the thesis of the symmetrical nature of land and other inputs. This in turn favoured the extension of the substitutability between land and complex labour from agricultural production to all kinds of production, including those in which land does not figure as a direct input. It so happened that whereas in classical economics the substitutability between land and complex labour presupposes that simple labour and equipment are strictly complementary, in neoclassical economics this substitutability is applied to all inputs indiscriminately.

However, the neoclassical interpretation of the law poses serious problems. In the first place, there is the problem of justifying, on empirical grounds, the general applicability of the substitution principle. Secondly, and more importantly, in order to allow the substitution of inputs to take place, a certain lapse of time is required during which the required modifications to the productive structure can be made. (It is certainly true that coal can replace oil to provide heating, but before this can happen it will be necessary to change the heating system.) The well-known distinction between the short run and the long run is a partial and indirect way to take the temporal element into consideration. In the short run the plant is fixed by definition. It is therefore the fixed input which, in the neoclassical interpretation of the law, plays the

same role as land in the classical interpretation. Now, neoclassical theory correctly states the law of diminishing returns with respect to the short run; however it is in the long run that the substitutability of inputs becomes actually feasible. One is therefore confronted with a dilemma: the neoclassical interpretation of the law seems to be more plausible in a long-run framework when there exists the necessary time to accommodate input adjustments; on the other hand, fixed inputs cannot, by definition, exist in the long run so that the law of variable proportions cannot be stated in such a context.

This dilemma is the price neoclassical theory has to pay for its interpretation of the law in accordance with the positivistic statute. Indeed, the power of deductive, truth-preserving rules of scientific inference is not purchased without a cost. A school of economic thought which is not prepared to sustain such a cost is the neo- Austrian. The neo-Austrian economists solve what has been called the fundamental tension by arguing economics cannot and should not provide general laws since, by its very nature, it is an idiographic and not a nomothetical discipline. The general target of economics is 'understanding' grounded in *Verstehen* doctrine: by introspection and empathy, the study of the economic process should aim at explaining individual occurrences, not abstract classes of phenomena. It follows that if by a scientific law one should mean a universal conditional statement of type 'for all x, if x is A, then x is B', statements regarding unique events cannot by definition express any regularity for the simple reason that any regularity presupposes the recurrence of what is defined as regular. In the words of L. von Mises, who shares with F. von Hayek the paternity of the neo-Austrian school, what assigns economics its peculiar and unique position in the orbit of pure knowledge '. . . is the fact that its particular theorems are not open to any verification or falsification on the ground of experience . . . the ultimate yardstick of an economic theorem's correctness or incorrectness is solely reason unaided by experience' (von Mises 1949, p. 858).

There is indeed a place for economic 'laws' in the framework of Austrian economics. The familiar 'laws' of economics (diminishing marginal utility, supply and demand, diminishing returns to factors, Say's Law and so on) are seen as 'necessary truths' which explain the essential structure of the economic world but with no predictive worth. In other words, economic laws are not generalizations from experience, as it is the case within the positivistic paradigm, but are theorems which enable us to understand the economic world. It is ironic that Mises' position of radical apriorism joined to Hayek's attack on scientism and methodological monism are completely at variance with the position taken by the father of the Austrian school, Carl Menger (1883), who announced that in economic theories exact laws are defined which are just as rigorous as in fact are the laws of nature.

Between the extreme positions of neoclassical positive economic and neo- Austrian economics are those who, without denying that economics is in search for laws in the same sense in which natural sciences are and that laws perform an explanatory as well as a predictive function, underline that the explicative structure of economics, albeit nomothetical, substantially differs from that of natural sciences. This intermediate position can be traced back to Keynes's (1973) methodology which considers the conditions of truth and universality of the positivistic conception of scientific laws as far too rigid for a discipline such as economics. Two main reasons account for the different epistemological status of laws in natural sciences and in economics. First, the knowledge of economic phenomena is itself an economic variable, that is, it changes, along with the process of its own acquisition, the economic situation to which it refers. The formulation of a new physical law does not change the course of physical processes; it does not influence the truth or falsity of the prognosis. This is not the case in economics where the prognosis, say, that in two years time there will be a boom can cause overproduction and a resulting recession. In turn, this specific aspect is strictly connected to the fact that the object of study of economics possesses an historical dimension. Economics is in time in a way that natural sciences are not. The ensuing mutability of observed regularities is well expressed by Keynes

when he writes, 'As against Robbins, economics is essentially a moral science and not a natural science. That is to say it employs introspection and judgements of value' (1973, p. 297) to which he adds, 'It deals with motives, expectations, psychological uncertainties. One has to be constantly on guard against treating the material as constant and homogenous' (p. 300).

Second, the role played by *ceteris paribus* clauses in natural sciences and in economics is substantially different. The modern economists appeal to the 'other things being equal' clause – which according to Marshall is invariably attached to any economic law – in all those cases where the classical economists were talking of 'disturbing causes'. J.S. Mill's (1836) discussion of inexact sciences is suggestive here:

> When the principles of Political Economy are to be applied to a particular case then it is necessary to take into account all the individual circumstances of that case . . . These circumstances have been called *disturbing causes*. This constitutes the only uncertainty of Political Economy. (1836, p. 300)

Also in natural sciences we find *ceteris paribus* clauses. Indeed, a scientific theory that could dispense with them would in effect achieve perfect closure, which is a rarity. So where lies the difference? The example of the science of tides used by Mill is revealing. Physicists know the laws of the greater causes (the gravitational pull of the moon) but do not know the laws of the minor causes (the configuration of the sea bottom). The 'other things' which scientists hold equal are the lesser causes. So could we conclude that just about all generalizations in both natural sciences and economics express in fact *tendency laws*, in the sense that these 'laws' truly capture only the functioning of 'greater causes' within some domain? Certainly not, since there is a world of difference between the two cases. Galileo's law of falling bodies certainly presupposes a *ceteris paribus* clause, so much so that he had to employ the idealization of a 'perfect vacuum' to get rid of the resistance of air. However, he was able to give estimates of the magnitudes of the amount of distortion that friction and the other 'accidents' would determine and which the law ignored. In other words, whereas in natural sciences the 'disturbing causes' have their own laws, this is not the case in economics where we find tendency statements with unspecified *ceteris paribus* clauses or, if specified, specified only in qualitative terms. In economics it is generally impossible to list all the conceivable inferences implied in a lawlike statement and to replace the *ceteris paribus* clause with precise conditions. So, for example, the law that 'less will be bought at a higher price' is not refuted by panic buying, nor is it confirmed by organized consumer boycotts. No test is decisive unless *ceteris* are really *paribus*.

These remarks help to understand the role acknowledged by Keynes to laws in economic inquiry. Besides general laws, there are also rules and norms which are significant in the explanation of economic behaviour. To Keynes, it makes no sense to reduce all forms of explanation in economics to that of the covering-law model. Indeed, whereas to justify a law one has to show that it is logically derivable from some other more general statements, often called principles or postulates, the justification of rules occurs through the reference to goals and the justification of norms through the reference to values which are not general sentences, but rather intended singular patterns or even ideal entities. Since no scientific law, in the natural scientific sense, has been established in economics, on which economists can base predictions, what are used and have to be used to explain or to predict are tendencies or patterns expressed in empirical or historical generalizations of less than universal validity, restricted by local and temporal limits. Recently, Arrow has amazed orthodox economists when raising doubts about the mechanistically inspired understanding of economic processes: 'Is economics a subject like physics, true for all time or are its laws historically conditioned?' (Arrow 1985, p. 322).

The list of generally accepted economic laws seems to be shrinking. The term itself has come to acquire a somewhat old-fashioned ring and economists now prefer to present their most cherished general statements as theorems or propositions rather than laws. This is no doubt a healthy reaction: for too long economists have been under the nomological prejudice, of positivistic origin, that

the only route towards explanation and prediction is the one paved with laws, and laws as forceful as Newton's laws. Images in science are never innocent: wrong images can have disastrous effects.

Bibliography

Arrow, K. 1985. Economic history: A necessary though not sufficient condition for an economist. *American Economic Review: Papers and Proceedings* 75: 320–323.

Keynes, J.M. 1973. The general theory and after. Part II: Defence and development. In *The collected writings of John Maynard Keynes*, vol. 14. London: Macmillan.

Menger, C. 1883. *Unterschungen über die Methode der Sozialwissenschaften*. Leipzig: Duncker & Humblot.

Mill, J.S. 1836. On the definition of political economy and the method of investigation proper to it. Repr. In: *Collected works of John Stuart Mill*, Essays on economy and society, vol. 4, ed. J.M. Robson. Toronto: University of Toronto Press, 1967.

von Mises, L. 1949. *Human action: A treatise on economics*. London: William Hodge.

Economic Man

Shaun Hargreaves-Heap and Colin G. Clark

Abstract

Economic man 'knows the price of everything and the value of nothing', so said because he or she calculates and then acts so as to satisfy best his or her preferences. The value of these preferences is immaterial. The hypothesis has nevertheless proved remarkably powerful not only in economics but across the social sciences where it has spawned 'rational choice' accounts of many aspects of social life. This ambition has attracted critics both from without and within. The latter have developed, with insights from psychology on how people acquire and use information, a less elegant but arguably more realistic model.

Keywords

Behavioural economics; Bounded rationality; Comparative statics; Economic man; Expected utility maximization; Game theory; *Homo economicus*; Hume, D.; Information economics; Institutional economics; Kant, I.; Law of small numbers; Learning; Neoclassical economics; Preferences; Rational behaviour; Rational expectations; Reference dependence; Risk; Satisficing; Self-serving biases; Social norms; Subjective probability; Welfare economics; White noise

JEL Classifications

B4

Among the many different portrayals of economic agents, the title of *homo economicus* is usually reserved for those who are rational in an instrumental sense. For example, this is how agency is defined in neoclassical economics. In its ideal type case the agent has complete, fully ordered preferences (defined over the domain of the consequences of his or her feasible actions), perfect information and all the necessary computing power. After deliberation, he or she chooses the action that satisfies their preferences better (or at least no worse) than any other. No questions are raised about the source or worth of preferences, reason focuses on the efficient selection of the means to given ends.

This basic model is then made more sophisticated. The theory of risk allows for the point that an action may have several possible consequences. When preferences are represented via the device of a utility function, the agent assesses his or her expected utility by discounting the utility of each consequence by how likely it is to be the actual one. That requires the agent to have a probability distribution for the consequences, even if only a subjective one. Other refinements include allowance for costs of acquiring information, of processing it and of action. Then there are complexities, illustrated by game theory, when actions of other agents form part of the environment in which the person acts. The basic vision remains, however, one of agents who are rational in the sense that they maximize an objective function subject to constraints (or act 'as if' this were the case).

This vision is not unique to neoclassical economics. For example, Marx's profit- maximizing capitalist fits the same instrumental model of rationality. Institutionalist accounts of, for instance, banks or trade unions often conceive economic bodies as similar unitary rational agents. Nor is the vision confined to any specific motivating desire in agents, like a selfish pleasure-maximizing drive. There is scope for allowing ethical preferences alongside the symptomatic textbook desires for apples and oranges. Agents are, however, regarded as self-interested, in the looser sense that they are moved to satisfy whatever preferences they happen to have. Furthermore, granted that *de gustibus non est disputandum,* this modest base is enough to ground a full-blown social theory on a model of agency which can be exported to other social sciences.

Such a social theory is individualist and contractarian, with a pedigree that includes Hobbes's *Leviathan* and Benthamite utilitarianism. The satisfaction of individual preference, aided by felicific calculation, is what makes the social world go round. Social relations become instrumental, in the sense that they embody exchanges in the service of individual preferences (see Becker 1976). For instance, marriage has been analysed in this spirit as an arrangement to secure the mutual benefit of exchange between two agents with different endowments. Crime has been claimed to occur because calculation of costs and benefits proves it to be the action that maximizes expected utility. Meanwhile, institutions, which feature in elementary microeconomics as constraints on individual choice, become deposits left by earlier transactions, often deliberately so as devices to prevent preferences being frustrated by situations of the Prisoner's Dilemma type. Government policies are explained on the hypothesis that the political arena is also peopled by individuals maximizing expected utility, who form coalitions in support of policies that will secure reelection (see Downs 1957). In short, *homo economicus* morphs into a universal *homo sapiens*.

Such a full-blown social theory may be too ambitious because assumptions that are plausible for simple market transactions become suspect when scaled up. For example, the ideal-type case makes agents, so to speak, transparent to themselves, and does not allow for history occurring behind their backs. Freudians would object to transparency of preferences and Marxians would invoke theories of false consciousness. (Although Marx's capitalists are instrumentally rational, their desire to maximize profit is an alienated one, 'forced' on them by a competitive capitalist system.) Many other social theorists would object to the treatment of norms and social relations as instrumental, on the grounds that norms are prior to preferences. For instance, cultural forms like the rules of orchestral composition are a source of musical preferences rather than a solution to a priori problems of maximizing musical enjoyment. Or, to put this differently, game theory yields too many instances of indeterminacy for an ambitious programme of reducing all social practices to the exercise of instrumental reason by the individual participating agents.

Such objections, of course, need not affect the more modest enterprise of explaining economic transactions within the parameters of social institutions like the market. But even here *homo economicus* has critics. Philosophically, it is not plain that preferences can be taken as given in a sense which makes them impervious to the agent's beliefs about the moral quality of his or her actions. In supposing that only desires can motivate agents, the economist is taking sides in a continuing philosophical dispute between Humeans, who regard reason as the slave of the passions, and Kantians, who make place for the rational monitoring of desire. This dispute surfaces plainly in welfare economics, when it is asked whether all preferences should count equally or whether 'capabilities' are more appropriate for the evaluation of social states than degrees of preference satisfaction, but bears on the elementary model of action too (see Sen 1999).

There are also methodological doubts about the empirical standing of the model. What would falsify the claim that economic agents seek the most effective means to satisfy their preferences? Apparent counter-examples can always be dealt with by treating them as evidence that preferences

have changed or been dismissed through a careful individuation of outcomes. Indeed, since preferences are unobservable, they can be identified only if the correctness of the model is presupposed. In other words, there is room for deeper dispute about the foundations of orthodox microeconomics than is always realized.

Even within economics there are critics. The most substantial attack comes from those who think that perfect information is not a useful limiting case of imperfect information. Granted that there is often no way of calculating the likely marginal costs and benefits of acquiring extra information (short of actually acquiring it), how shall the agent decide rationally when to stop? Simon (1976) uses the question to argue for 'satisficing' models, in place of maximizing ones, and for 'procedural' or 'bounded' rationality. Rationality, he suggests, is a matter of following a procedure that halts with a good solution, and should not be defined in terms of best solutions. While this is a tempting thought, it is not obvious that searching for a 'good' solution is any easier than the best one if 'good' is some kind of second-best version of the 'best'. As a result, 'behavioural economists' have been drawn to the large experimental literature in psychology on how people actually behave and have produced economic models of decision-making that incorporate a variety of psychological processes such as 'self-serving biases', the 'law of small numbers' and 'reference dependence' (see Kahneman 2003). In this way, *homo economicus* has become more psychologically complex and more of an institutional or organizational person than an abstract maximizer.

The rational expectations hypothesis offers a different approach to the information issue. A rational agent who is short of information should not use an information-generating mechanism that gives rise to systematic errors. If errors are systematic, the agent should be able to learn how to eliminate them by amending the mechanism. There is an incentive to do so, because improved estimates of future variables will be profitable. On the face of it this makes rational expectations the natural ally of the pure economic-man models. Economic Man can proceed much as before, in the assurance that inadequate information involves nothing more systematic than 'white noise' and with the benefit of fresh analytic results that flow from a rational expectations hypothesis.

But this is to sidestep the informational problem set earlier, unless one sees how rational agents will learn to remove systematic errors. When there are costs to learning then it may not be rational to expend the effort that achieves a rational expectation. If we set such costs aside, in some simple learning situations a Bayesian updating procedure turns a rational expectations-generating process into an approximation of adaptive expectations, which could be construed as a procedural rule of thumb. But no general rapprochement between maximizing and procedural models of rationality follows. In more general learning situations the rational agent is trying to learn the rational expectations equilibrium relationship between variables – the one which, if used by agents to form their expectations, would reproduce itself in experience (white noise apart). This sounds easy, in that repeated experience of a particular relationship should lead to convergence on accurate parameter estimates. However, ignorance of the rational expectations equilibrium values produces behaviour that departs from those values. So observed values of variables embody a distortion which agents cannot correct without knowing the dimensions of their own ignorance. To know this, however, they would have to know the rational expectations equilibrium values already. To put it as the procedural critics might, learning would be feasible only if there were nothing to learn. The information question has been begged; and the door again opens on to psychology and its rich literature on what people actually do.

Nevertheless, the ideal-type Economic Man remains a powerful model of action not only in neoclassical theories, where insights in comparative statics have been especially notable, but elsewhere too. How powerful it finally is depends, within economics, on what becomes of the informational difficulties and on whether procedural or bounded models can come up with rival results of equal scope and elegance. For the wider social sciences, it offers a tempting analysis of social

behaviour at large both for transactions in other social arenas and for the emergence of the institutions that govern those arenas. But the greater its ambitions, the more serious become the unresolved doubts about the origin of preferences and their relation to norms and institutions.

See Also

- ▶ Altruism, History of the Concept
- ▶ Rational Behaviour
- ▶ Rationality, History of the Concept
- ▶ Utilitarianism and Economic Theory

Bibliography

Becker, G. 1976. *The economic approach to human behaviour*. Chicago: Chicago University Press.

Downs, A. 1957. *An economic theory of democracy*. New York: Harper Row.

Kahneman, D. 2003. Maps of bounded rationality: Psychology for behavioural economics. *American Economic Review* 93: 1449–1475.

Sen, A. 1999. *Commodities and capabilities*. Oxford: Oxford University Press.

Simon, H.A. 1976. From substantive to procedural rationality. In *Method and appraisal in economics*, ed. S. Latsis. Cambridge: Cambridge University Press.

Economic Organization and Transaction Costs

Steven N. S. Cheung

One important extension of the Coase Theorem states that, if all costs of transactions are zero, the use of resources will be similar no matter how production and exchange activities are arranged. This implies that in the absence of transaction costs, alternative institutional or organizational arrangements would provide no basis for choice and hence could not be interpreted by economic theory. Not only would economic organization be randomly determined; there actually would not be any organization to speak of: production and exchange activities would simply be guided by the invisible hand of the market.

But organizations or various institutional arrangements do exist, and to interpret both their presence and their variation, they must be treated as the results of choice subject to the constraints of transaction costs.

In the broadest sense transaction costs encompass all those costs that cannot be conceived to exist in a Robinson Crusoe economy where neither property rights, nor transactions, nor any kind of economic organization can be found. This breadth of definition is necessary because it is often impossible to separate the different types of cost. So defined, transaction costs may then be viewed as a spectrum of institutional costs including those of information, of negotiation, of drawing up and enforcing contracts, of delineating and policing property rights, of monitoring performance, and of changing institutional arrangements. In short, they comprise all those costs not directly incurred in the physical process of production. Apparently these costs are weighty indeed, and to term them 'transaction costs' may be misleading because they may loom large even in an economy where market transactions are suppressed, as in a communist state.

By definition, an organization requires someone to organize it. In the broadest sense, all production and exchange activities not guided by the invisible hand of the market are organized activities. Thus, any arrangement that requires the use of a manager, a director, a superviser, a clerk, an enforcer, a lawyer, a judge, an agent, or even a middleman implies the presence of an organization. These professions would not exist in the Crusoe economy, and payments for their employment are transaction costs.

When transaction costs are defined to include all costs not found in a Crusoe economy, and economic organizations are defined equally broadly to include any arrangement requiring the service of a visible hand, a corollary appears: all organization costs are transaction costs, and vice versa. That is why during the past two decades economists have striven to interpret the various forms of organizational arrangements in terms of the varying costs of transactions.

Some obvious examples will illustrate the point. A worker in a factory (an organization) may be paid by a piece rate or by a wage rate. If the costs of measuring and enforcing performance (one type of transaction cost) are zero, then either arrangement will yield the same result. But if these costs are positive, the piece-rate contract will more likely prevail if the costs of measuring outputs are relatively low, whereas the wage contract will more likely be chosen if the costs of measuring hours and enforcing performance are low relative to the costs of measuring outputs. As another example, some restaurants (again an organization) measure the quantity of food sold; others serve buffet dinners, allowing customers to eat as much as they please at a fixed price per head. The cost of metering and quantifying food consumption relative to the basic cost of the food will determine which arrangement is chosen. In the total absence of transaction costs, the factory or the restaurant would not exist in the first place, because consumers would buy directly from the input owners who produce the goods and services.

As early as 1937, R.H. Coase interpreted the emergence of the firm (an organization) in light of the costs of determining market prices (transaction costs). When these costs are substantial because of the difficulties of measuring separate contributions by workers and of negotiating prices for separate components of a product, a worker may choose to work in a factory (a firm); he surrenders the right to use his labour by contract and voluntarily submits to direction by a visible hand, instead of personally selling his services or contributions to customers through the invisible hand of the market. The firm is therefore said to supersede the market. As the supersession progresses, the saving in the costs of determining prices will be countered by the rising costs of supervision and of management in the firm. Equilibrium is reached when, at the margin, the cost saving in the former equals the rising cost in the latter.

The firm superseding the market may be regarded as a factor market superseding a product market. If all costs of transactions were zero, the two markets would be inseparable in that a payment made by a customer to the owner of a factor of production would be the same as payment made to a product seller. In such a world it would be a fallacy to speak of the factor market and the product market as coexisting entities.

The presence of transaction costs is a prelude to separate the factor market from the product market. However, in some arrangements, such as the use of certain piece rates, it may become impossible to separate the one market from the other. Therefore, instead of viewing the firm as superseding the market, or the factor market as superseding the product market, it is more correct to view the organizational choice as one type of contract superseding another type. In these terms, the choice of organizational arrangements is actually the choice of contractual arrangements.

When organizational choices are viewed as contractual choices, it becomes evident that it is often impossible to draw a clear dividing line separating one organization from another. Take the firm, for example. It is often the case that the entrepreneur who holds employment contracts (and it is not clear whether it is the entrepreneur who employs the workers or the workers who employ the entrepreneur) may contract with other firms; a contractor may subcontract; a subcontractor may sub-subcontract further; and a worker may contract with a number of ‘employers’ or ‘firms’. If the chain of contracts were allowed to spread, the ‘firm’ might encompass the whole economy. With this approach the size of the firm becomes indeterminate and unimportant. What are important are the choice of contracts and the costs of transactions that determine this choice.

Traditional economic analysis has been confined to resource allocation and income distribution. Contractual arrangements as a class of observations have been slighted in that tradition. In a world complicated by transaction costs, this neglect not only leaves numerous interesting observations unexplained, but actually obscures the understanding of resource allocation and income distribution. The economics of organization or institution or, for that matter, the workings of various economic systems, were never placed in the proper perspectives under the traditional approach. For generations students were told that

various kinds of 'imperfections' were the cause of seemingly mysterious observations: policies were 'misguided', or antitrust specialists were barking up the wrong trees.

The costs of introducing new and more valid ideas must have been enormous. Even today textbooks still discuss marginal productivity theory only with reference to fixed wage and rental payments. Yet economists have known all along that (for labour alone) payments may be in the peripheral forms of piece rates, bonuses, tips, commissions, or various sharing arrangements; moreover, even wage rates may assume a number of forms. Each type of contract implies different costs of supervision, of measurement, and of negotiation, and the form of economic organization, along with the function of the visible hand, changes whenever a different contractual arrangement is chosen.

The choice of contractual arrangements is not, of course, confined to the factor markets. In the product markets, pricing arrangements such as tie-in sales, full-line forcing, or membership fees associated with clubs, may similarly be interpreted in light of transaction costs. Further, business organizations in mergers, franchises, and various forms of integration are now beginning to be viewed as transaction-cost phenomena. Indeed, close inspection of department stores and shopping centres reveals pricing and contractual arrangements between a central agent and individual sellers, as well as among the sellers themselves, which could not be explained by textbook economics.

Transaction costs are often difficult to measure and, as noted earlier, difficult to separate by type. However, the measurement problem can be avoided if only we are able to specify how these costs vary under different observable circumstances, and their different types are separable if viewed in terms of changes at the margin. These two conditions are requisite in the derivation of testable implications for the interpretation of organizational behaviour.

The use of transaction costs to analyse institutional (organizational) choice is superior to three other approaches. One approach would focus on incentives. However, incentives are not in principle observable, and we will do better in deriving testable propositions if the same problem is viewed in terms of the costs of enforcing performance. A second approach adopts risk. However, it is difficult to ascertain how risk is altered under different circumstances. Many risk problems, such as the uncertainty of whether an agreement will be honoured, are also problems of transaction costs, and it is easier to deal directly with those. Finally, some recent advances in transaction-cost analysis have called attention to the costs embodied in dishonesty, cheating, shirking, and opportunistic behaviour. Yet these are loose terms and, whatever they describe, to some extent are always to be found. To the degree that we can identify the particular costs of transactions that promote dishonesty, that shadowy explanation is no longer needed. After all, in what sense can we say a person is 'increasingly dishonest' or 'increasingly opportunistic'?

The transaction-cost approach to analysis of economic organizations can be extended upward from a few participants to the 'government' or even the nation itself. At the lower level, the owners of condominium units almost as a rule form associations with specific by-laws and elect committees to act on matters of common concern, the decisions being determined by majority vote. The transaction costs of ballot voting are less than those of using prices and dollar votes in certain circumstances, and trivial matters may even be delegated to a 'dictatorial' manager to further reduce the cost of voting. Similarly, residents in a particular location may choose to incorporate into a city, selecting their own mayor, with a committee setting up the building codes, hiring firemen and policemen, and deciding other matters of common concern.

Private property rights offer the unique advantage of allowing individual property owners the option of *not* joining an organization. This choice is an effective restraint against the adoption of an organization with higher transaction costs. It is true that a home-owner in a given region may, by majority vote, lose his option of not joining in a city corporation (unlike a worker who, in a free enterprise economy, always has the option of not joining a 'firm'). But with private property

rights the majority vote aims at cost saving, and a reluctant resident may exercise his own judgement by selling his house and moving elsewhere.

Private property rights further reduce transaction costs under competition. An entrepreneur or agent who wants to recruit other resource owners to join his organization must, under competition, offer attractive terms, and this can be achieved only if his organization can effectively reduce transaction costs. On the other hand, the resource owner competing to join an organization will be more inclined to deliver a good performance when at risk of losing his job.

The option of not joining an organization and the cost-reducing function of competition are, of course, restrained when an organization is extended to encompass an entire nation. When citizenship is dictated by birth, the option of not joining is restrained, and competition among nations to recruit members is decidedly less than among organizations within a nation. This relative lack of cost-reducing mechanisms is all the more evident in a communist state, where a citizen does not have the option of choosing an organization within that state.

A communist state may be regarded as a 'superfirm' in which comrades lack the option of not joining. Each worker is assigned to a particular job supervised and directed by the visible hands of comrade officials of varying ranks. In this aspect the communist state is remarkably similar to what Coase calls a 'firm', where workers are told what to do instead of being directed by market prices. But the lack of market prices in the communist state is not due to the costs of determining prices; rather, in the absence of private property rights market prices simply do not exist, and visible supervision by a hierarchy ranking becomes the remaining alternative to chaos.

The transaction costs of operating an organization are necessarily higher in a communist state than in a free enterprise economy, due to the lack of option of not joining and the lack of competition both to recruit members among organizations and to induce members to perform well.

If the transaction costs of operating organization were zero, resource allocation and income distribution would be the same in a communist state as in a free enterprise state: consumer preferences would be revealed without cost; auctioneers and monitors would provide freely all the services of gathering and collating information; workers and other factors of production would be directed free of cost to produce in perfect accord with consumer preference; each consumer would receive goods and services in conformity with his preferences; and the total income received by each worker, as determined costlessly by an arbitrator, would equal his marginal productivity plus a share of the rents of all resources other than labour, according to any of a number of criteria costlessly agreed upon. But such an ideal situation is obviously not to be found.

We therefore conclude that the poor economic performance of a communist state is attributable to the high transaction costs of operating that organization. Under the postulate of constrained maximization, the communist state survives for the same reason that any 'inefficient' organization survives: namely, the transaction costs of *changing* an organizational (institutional) arrangement are prohibitive. Such costs include those of obtaining information about the workings of alternative institutions, and of using persuasive or coercive power to alter the status of the privileged groups whose incomes might be adversely affected by the institution of a different form of economic organization.

See Also

- ▶ Coase Theorem
- ▶ Vertical Integration

Bibliography

Alchian, A.A., and H. Demsetz. 1972. Production, information costs, and economic organization. *American Economic Review* 62: 777–795.

Barzel, Y. 1982. Measurement costs and the organization of markets. *Journal of Law and Economics* 25: 27–48.

Cheung, S.N.S. 1969. Transaction costs, risk aversion, and the choice of contractual arrangements. *Journal of Law and Economics* 12: 23–42.

Cheung, S.N.S. 1982. *Will China go 'capitalist'?* Hobart Paper 94. London: International Economic Association.
Cheung, S.N.S. 1983. The contractual nature of the firm. *Journal of Law and Economics* 26: 1–21.
Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.
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.
Klein, B., R.G. Crawford, and A.A. Alchian. 1978. Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.
Knight, F.H. 1921. *Risk uncertainty and profit*. Boston: Houghton Mifflin.
McManus, J.C. 1975. The costs of alternative economic organizations. *Canadian Journal of Economics* 8: 334.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and anti-trust implications*. Glencoe: Free Press.

Economic Sanctions

Jeffrey J. Schott

Abstract

Economic sanctions are tools of statecraft used to achieve a broad range of foreign policy goals by threat or deployment of coercive measures such as trade embargoes, asset freezes, or withholding of development aid. Throughout the post-war era, the United States and other countries frequently have imposed economic sanctions, even though they have contributed only infrequently to foreign policy successes. Globalization has made the exercise of economic coercion increasingly complex, but has not obviated the utility of sanctions as part of the foreign policy arsenal.

Keywords

Economic development; Economic sanctions; Globalization; Terrorism

JEL Classifications

F51

Economic sanctions are tools of statecraft used to influence the behaviour of foreign countries by the threat or actual withdrawal of trade and sources of finance. Traditional means of coercion include trade embargoes, withholding development assistance, and asset freezes. The objective is to confront a foreign country with a choice: either bear the cost of lost trade and finance, or change policies to comply with the demands of those imposing the sanctions (the sender countries). Projecting power through economic coercion is deemed more forceful than diplomatic reproach yet less drastic than military intervention. In practice, economic measures generally are deployed as part of a broader programme of foreign policy responses encompassing diplomatic entreaties, covert or quasi-military intrusions, and threat of or preparation for military action.

Countries impose sanctions in pursuit of a variety of foreign policy goals. Historically, economic sanctions have preceded and then accompanied military conflict. The oil embargo of Japan was a prelude to the Second World War in the Pacific; so, too, were the United Nations' sanctions against Iraq following its invasion of Kuwait in 1990. Obviously, sanctions are part and parcel of 'hot wars' that sever economic ties between the combatants; but they are also prevalent in 'cold war' episodes, where the goal is to impair military capabilities through denial of weapons and dual-use technologies (for example, post-war sanctions against the Soviet Union and its satellites under the auspices of the Consultative Group and Coordinating Committee for Multilateral Export Controls, or CoCom, and efforts to blunt the development of nuclear weapons in Iran and North Korea). In addition, sanctions have sought to impede or reverse military incursions across borders (for example, the League of Nations effort to get Italy to withdraw from Abyssinia in 1936) and between warring factions within a country (the sad recent history of several West African states).

Not all sanctions episodes respond to or presage military actions. Many post-war cases have been advanced to counter other types of aberrant behaviour such as state sponsored terrorism, proliferation of weapons of mass destruction, or

human rights abuses. In these cases, sender countries impose sanctions in an effort to redress foreign outrages, to deter emulation by others (the rationale in most anti-proliferation cases), and to punish the target regime for its misdeeds (for example, the US grain embargo after the Soviet invasion of Afghanistan in 1989). In a number of cases, sanctions pursue the goal of regime change *sotto voce* – whether the target is Moammar Gaddafi in Libya, Kim Jong-il in North Korea, or the Afrikaaners in South Africa. Sanctions that portend regime change obviously meet stauncher resistance than those that seek narrow changes in governance by the target government.

Do Sanctions 'Work'?

Foreign policy ventures seldom yield unambiguous results. Gauging the effectiveness of sanctions involves a combination of quantitative method and intuition, and often requires subjective evaluation of incomplete results. Sanctions alone seldom are sufficient to change foreign practices, but they can *contribute* to the achievement of policy goals in conjunction with other instruments of statecraft, if properly designed and implemented. That is easier said than done.

Sanctions are blunt policy instruments; they are better at impairing economic performance over time than at inflicting surgical strikes on target countries. Senders that expect immediate gratification often tire of the effort, especially if the sanctions impose significant costs on their own firms and workers. Moreover, when sanctions are hard hitting, it is difficult to avoid innocent victims within the target country and in neighbouring states; in such cases, the debilitating effect of sanctions often results in substantial suffering among the civilian population. Humanitarian exemptions from the sanctions designed to soften the blow to the general public invariably weaken the economic impact of the sanctions and muddy the policy signal to the target regime. To be sure, such loopholes in the sanctions net are important both on moral grounds and to maintain the cohesion of the coalition of sender countries, but the loopholes are prone to abuse (witness the scandalous operation of the United Nations' oil-for-food programme, which was supposed to channel Iraqi oil export revenues to humanitarian assistance) and reduce the economic pressure to comply with the sender's demands.

Almost all sanctions leak; targeted countries can evade the full thrust of the economic restrictions by redirecting trade and finance to non-sanctioning states or by engaging in clandestine operations. Countries seeking economic or political influence with the target regime often conspire to evade the sanctions; the Cold War period was replete with examples of 'Black Knight' countries coming to the rescue of targeted regimes with aid to offset the impact of sanctions imposed by the United States or the Soviet Union. Smugglers still outwit even the most comprehensive embargoes – witness the billions of dollars earned by Saddam Hussein, the former Iraqi president, from illicit oil exports during the period of 'comprehensive' UN sanctions against Iraq. For a price, targeted regimes can still procure goods, services and technologies; the profit motive seems to be an irresistible force regardless of region or culture!

That said, sanctions have contributed to a few notable successes in the post-war era, including the collapse of the apartheid regime in South Africa and the renunciation of terrorism by President Gaddafi in Libya. Hufbauer et al. (2007) found success – measured by the partial fulfillment or better of policy goals – in more than a quarter of the almost 200 sanctions episodes documented in the 20th century. (The third edition of this comprehensive study of economic sanctions contains updated policy analysis and case studies, and an extensive bibliography. See also Baldwin 1985, for an examination of the tools of economic statecraft, and Martin 1992, for analysis of the use of multilateral economic sanctions.) Most of these cases, however, involved relatively modest demands on the target country. When the stakes are high, resistance by the target regime stiffens. Accordingly, most high-profile sanctions cases – like those seeking to oust President Castro in Cuba or to deter support for terrorism and the development of nuclear weapons by the ayatollahs in Iran – have been abject failures.

Can Sanctions Be Effective in an Era of Rampant Globalization?

Economic sanctions traditionally have been the domain of big powers, acting unilaterally or as part of a broader international coalition. Until recently, the big powers controlled the trade lanes and purse strings of international commerce, and held a near monopoly on advanced technologies. Since the mid-1980s, however, the success of post-war economic development, spurred in part by the spread of technological innovation, has eroded the franchise of the big powers and created alternative sources of goods, technology and capital for countries targeted by economic sanctions. Simply put, globalization has made it much harder to design an effective sanctions policy.

In addition, global politics are now more complex than in the period of East–West rivalry. Former allies differ regarding strategies and priorities for using sanctions to deal with regional trouble spots. For example, Europe is more vulnerable than the United States to an interruption of energy supplies from the Middle East, and thus is less willing to constrain oilfield development and to take actions that risk political retaliation. Similarly, China and Japan are highly dependent on imported energy and thus sensitive to sanctions against Iran and other oil – producing states.

Globalization also has contributed to the decentralization of power, allowing smaller countries – especially those rich in energy resources – to provide offsetting assistance to blunt the economic impact of sanctions. But the influence of globalization goes beyond the realm of state-to-state intervention; terrorism, for example, now operates in a stateless domain of sleeper cells and territories outside of governmental control linked through informal financial and telecommunications networks. For that reason, sanctions policies increasingly seek to target individuals and corporations as well as governmental bodies, and to favour financial measures to interdict inter-bank electronic transfers in addition to the more traditional controls on trade, investment and development assistance.

In sum, economic sanctions continue to play a major role in international relations. However, the familiar goals of economic coercion now must be pursued through measures adapted to the changing conditions in global markets. The use of economic sanctions needs to be reconsidered and revamped, but not abandoned.

See Also

- ▶ Foreign Aid
- ▶ Trade Policy, Political Economy Of
- ▶ Transfer Of Technology

Bibliography

Baldwin, D.A. 1985. *Economic statecraft*. Princeton: Princeton University Press.

Hufbauer, G.C., K. Elliott, J.J. Schott, and B. Oegg. 2007. *Economic sanctions reconsidered*. 3rd ed. Washington, DC: Peterson Institute for International Economics.

Martin, L.L. 1992. *Coercive cooperation: Explaining multilateral economic sanctions*. Princeton: Princeton University Press.

Economic Science and Economics

Henry Sidgwick

The terms ‘economy’ and ‘economic’ or ‘economical’, are now used chiefly in two meanings, which it is well to distinguish clearly; since, though divergent in their history, they are liable to fusion, and therefore in some degree to confusion.

‘Economy’ originally meant, in Greek, the management of the affairs of a household, especially the provision and administration of its income. But since both in the acquisition and in the employment of wealth it is fundamentally important to avoid waste either of labour or of its produce, ‘economy’ in modern languages has come to denote generally the principle of seeking to attain, or the method of attaining, a desired end with the least possible expenditure of means; and the words ‘economy’, ‘economic’, ‘economical’, are often used in this

sense, even without any direct relation to the production, distribution, or consumption of wealth. Thus we speak of 'economy of force' in a mechanical arrangement without regard to its utility, and of 'economy of time' in any employment whether productive of wealth or not.

On the other hand, as there is an obvious analogy between the provision for the needs of a state and the provision for the needs of a household, 'political economy', in Greek, came to be recognized as an appropriate term for the financial branch of the art or business of government. It is found in this sense in a treatise translated as Aristotle's in the 13th century; and so, when, in the transition from medieval to modern history, the question of ways and means obtrusively claimed the attention of statesmen, 'political economy' was the name naturally given to that part of the art of government which had for its aim the replenishment of the public treasury, and – as a means to this – the enrichment of the community by a provident regulation of industry and trade. And the term retained this meaning till the latter part of the 18th century without perceptible change – except that, towards the end of this period, the enrichment of the people came to be less exclusively regarded from the point of view of public finance, and more sought as a condition of social well-being.

But in the latter part of the 18th century, under the influence primarily of the leading French 'Économistes' or 'Physiocrats' – Quesnay, De la Rivière, and others - the conception of political economy underwent a fundamental change, in consequence of a fundamental change in the kind of answer which these thinkers gave to the question 'how to make a nation wealthy'. The physiocrats proclaimed to France, and through France to the world, that a statesman's true business was not to *make* laws for industry and trade in the hope of increasing wealth; but merely to ascertain and protect from encroachment the simple and immutable laws of nature, under which the production of wealth would regulate itself in the best possible way if governments would abstain from meddling. A view broadly similar to this, but less extreme, and, partly for this reason, more directly influential, was expounded in Adam Smith's *Wealth of Nations*. Instead of showing the statesman how to 'provide a plentiful revenue or subsistence for the people' – which was one of the two main objects of political economy, according to the traditional view – Adam Smith aims at showing him how nature, duly left alone, tends in the main to attain this end better than the statesman can attain it by governmental interference. Accordingly, so far as the widespread influence of Adam Smith's teaching went, that branch of the statesman's art which aimed at 'providing a plentiful revenue for the people' tended almost – though not altogether – to shrink to the simple maxim of *laisser faire*: leaving in its place a scientific study of the processes by which wealth is produced, distributed, and exchanged, through the spontaneous and partly unconscious division of labour among the members of human society, independently of any governmental interference beyond what is required to exclude violence or fraud. A part, indeed, of the old art of political economy – that which aimed at 'supplying the state with a revenue sufficient for the public service' – remained indispensable to the statesman; but it was held that this traditional art required to be renovated by being rationally based on the doctrines of the new-born science just described. It is, then, this scientific study of the department of social activity that most writers on the subject now primarily mean by the term 'political economy': such part of the old governmental art so called, as the doctrine of the new science is held to admit, being commonly regarded as 'applied political economy'. In consequence of this change the adjective 'economic', instead of the too cumbrous 'politico-economic', has come to denote the matters investigated by the science of political economy, and the propositions and arguments relating to them.

By thinkers and duly-instructed students this distinction between 'science' and 'art' – between the study of 'what is' and the study of 'what ought to be' – is usually regarded as simple and clear; and accordingly when such persons speak of the 'laws of political economy' they mean not rules by which the process of the social production and distribution of wealth *ought* to be governed, but general relations of co-existence and sequence among phenomena of this class, ascertained by a scientific study of this process as it actually takes place.

This distinction, however, has been found difficult to establish in common thought: even well-educated persons still occasionally speak of the 'laws of political economy' as being 'violated' by the practice of statesmen, trades-unions, and other individuals and bodies. It is partly in order to prevent this confusion that the terms 'economic science' and 'economics' have recently come more and more into use, as a preferable alternative for political economy, so far as it is the name of a science. As to the scope of this science – it would be generally agreed that it is a branch of a larger science, dealing with man in his social relations; that it is to an important extent, but not altogether, capable of being usefully studied in separation from other branches of this science; and that it is mainly concerned with the social aspect – as distinct from the special technical aspect – of such human activities as are directed towards the production, appropriation, and application of the material means of satisfying human desires, so far as such means are capable of being exchanged. It would also be generally agreed that the method of economic science is partly deductive, partly inductive and historico-statistical. But to attempt a more precise determination of its method and scope, and especially of its relation to the art or system of practical rules which should guide the action of governments or private individuals in economic matters, would require us to enter into questions of a highly controversial kind; which will be more conveniently discussed when we come to deal with the older and wider term Political Economy.

Reprinted from *Palgrave's Dictionary of Political Economy.*

Economic Sociology

Richard Swedberg

Abstract

The term 'economic sociology', used primarily by sociologists, is defined as the application of sociological concepts and methods of analysis to economic phenomena. Founded by Durkheim, Weber, and Simmel, and continued by Schumpeter and Polanyi, it began to flourish in the mid-1980s around the notion that economic actions are embedded in personal networks. The concept of networks and other concepts and perspectives from 'new economic sociology' facilitate the analysis of topics like the links between corporations and between firms, job search, production markets, finance markets, insurance markets, industrial markets, consumption, and ethnic entrepreneurship. Its long-term impact on economics remains uncertain.

Keywords

Becker, G; Capitalism; Weber on; Polanyi on; Coase, R; Division of labour; Smith vs Durkheim on; Durkheim, E; Economic sociology; Emotions; Entrepreneurship; Ethnic; And immigration; Forced; Granovetter, M; Immigration; And ethnic entrepreneurship; Jevons, S; Marx, K; Networks; Strong vs weak ties; And job search; And consumption; And groups of firms; New Institutional Economics; Parsons, T; Polanyi, K; On capitalism; Schumpeter, J; On social economics;; Simmel, G; On trust; Smith, A; Trust; Simmel on; Weber, M; On capitalism

JEL Classifications

Z1

The first recorded use of the term 'economic sociology' is in a 1879 work by Stanley W. Jevons; and it is clear from the context that Jevons viewed economic sociology as part of the overall enterprise of economics rather than as an area belonging to another social science, such as sociology. Today, in contrast, the term 'economic sociology' is used primarily by sociologists, and they define it as *the application of sociological concepts and methods of analysis to economic phenomena.* While it is definitely possible to treat the great concern with institutions in New Institutional Economics, for example, as a kind of economic sociology, the reader is referred to the entry for

this topic for this type of analysis. Similarly, while Gary Becker at times has referred to his extension of the economic model to non-economic topics as 'economic sociology', the reader is similarly referred to the entry for his work.

Here, the first section, on classical economic sociology, is followed by sections on more recent economic sociology. This way of proceeding not only follows the general development of the field of economic sociology but is often how economic sociology is taught today, since the classics play a somewhat different role in economic sociology (as in sociology itself) to that in economics. In brief, while sociologists are trained through work with the classics as well as modern material, today's economists read the classics primarily when they study the history of their discipline.

Classical Economic Sociology

The work of Karl Marx (1818–83) can be seen as a type of economic sociology, in the sense just mentioned. More generally, Marx closely linked classical economic categories, such as value, price and capital, to distinctly social categories, such as class, work and relations of production. Nevertheless, Marx has played a marginal role in economic sociology as an academic enterprise – except as a catalyst and inspiration for a number of scholars, including Max Weber and Joseph Schumpeter.

Modern academic sociology is generally regarded as having three founders – Max Weber, Emile Durkheim and Georg Simmel – all of whom were interested in the economy. Georg Simmel (1858–1918), who pioneered sociology in Germany, wrote on the sociological role of money, competition and trust in the economy (Simmel 1900; 1908). He closely linked different types of money to different types of social authority, and also attempted to show how money is linked to the element of relativism in modern society. Competition, he argued, releases the energy of all participants to the benefit of the public, whereas in a conflict combatants are pitted against each other and block each other's efforts. Trust, finally, is central to the economy as well as society at large; without trust, the economy as well as society would collapse.

Emile Durkheim (1858–1917), unlike Simmel, attempted to institutionalize economic sociology, partly by encouraging some of his students to specialize in this field. Durkheim's own most important contribution to economic sociology can be found in his doctoral study of the division of labour, which contains a sharp critique of the argument in Adam Smith's *The Wealth of Nations* (Durkheim 1893). According to Durkheim, while Adam Smith had seen the significance of division of labour exclusively from the perspective of the creation of wealth, he had neglected its importance for the cohesion of society. More precisely, Smith had failed to realize that the primary function of the division of labour in modern society is to tie people together: people who do very different things need each other, and this is also what gives cohesion to modern society.

The most sustained effort to lay a solid theoretical foundation for economic sociology and also to carry out empirical studies can be found in the work of Max Weber (1864–1920) (Swedberg 1998). While Weber is famous for *The Protestant Ethic and the Spirit of Capitalism* (1905), it is less well known that his work is part of a more general attempt to develop a new academic field that would complement economic history and economic theory, namely, economic sociology.

At first Weber carried out empirical and historical studies with this goal in mind, and of these *The Protestant Ethic* is by far the best known (but see also Weber 1909; 1895). Weber's thesis, which holds that a certain type of religion ('ascetic Protestantism') had helped to create the mentality of modern capitalism in the 16th and 17th centuries ('rational capitalism'; Weber 1905), has led to a heated debate. Most commentators have found Weber's thesis unconvincing, but it should be emphasized that the debate is still going on with as much fervour as in the early 20th century (see, for example, Marshall 1982).

The heart of Weber's economic sociology is to be found in *Economy and Society*, a work that was incomplete when Weber died. It is here, for example, that Weber set out his well-known typology of

capitalism: political capitalism, traditional capitalism and rational capitalism. While the former two have existed for thousands of years, rational capitalism has emerged only in modern times and in the West. While traditional capitalism is non-dynamic and centred around small enterprises involving trade and the exchange of money, political capitalism is profit-making that either takes place through the state or under its direct protection, as in imperialism. Rational capitalism, in contrast, gets its name from the strong element of conscious and methodical calculation: the activities of the firm are carried out with the help of accountants and a trained staff; similarly, the activities of the state bureaucracy (including in the legal system) are predictable and rational. All of this makes possible a truly dynamic and revolutionary form of capitalism, according to Weber.

Economy and Society also contains a serious attempt by Weber to develop the central theoretical categories of economic sociology (Weber 1914, pp. 63–211). The basic unit of analysis is 'economic social action', which differs from economic action in economic theory by partly being determined by its social dimension. Economic social action is defined by Weber as behaviour that is (*a*) invested with meaning, (*b*) aimed at utility and (c) *oriented to another actor.* Utility is what makes the action 'economic'; and Weber's definition of 'social' is to be found in the formula 'orientation to another actor'. The emphasis on meaning explains why Weber's sociology is called an interpretive sociology; his economic sociology was to be a form of *interpretive economic sociology.*

Weber then proceeds to economic relationships in which two actors orient their actions to one another. These relationships can be either open or closed; and there is a general tendency for open economic relationships to become closed when there are not enough resources to go round. Economic organizations are defined as closed social relationships of a certain type; there also has to be a staff. Economic systems, finally, can be oriented either to profit-making (as in capitalism) or to the provision for a household (as in socialism or earlier non-market economies). Weber also discusses a host of other topics, including trade, money, division of labour and different ways of appropriation.

After the Classics

While the founding fathers of sociology were all interested in economic sociology and promoted it, the topic did not become popular among sociologists until the mid- 1980s with the emergence of so-called 'new economic sociology'. The reason for this is not clear, but may well have been a strong sense among sociologists that the economists were better equipped to deal with economic topics. In any case, very little work on economic sociology was produced between 1920 and the mid-1980s.

There were, however, a few exceptions. For one thing, sociologists did discuss topics relating to the economy, even if they did so under labels other than 'economic sociology'. One example is industrial sociology, which saw as its main task to analyse situations when people work in groups, in the factory as well as the office. An important research result is that workers develop norms in a number of areas, including what is seen as the maximum effort. Those who breach these norms are punished (for example, Whyte 1955).

Three individuals who all made important contributions to economic sociology also appeared during the period after the classics: Joseph Schumpeter, Karl Polanyi and Talcott Parsons. According to Schumpeter (1885–1950), economics should be a broad science ('social economics') and encompass four areas: economic theory, economic history, economic statistics and economic sociology (Schumpeter 1954, pp. 12–24). Schumpeter did work in each of these fields, including economic sociology. According to Schumpeter, economic sociology deals with institutions, while economic theory deals with economic mechanisms. Schumpeter's three most famous essays in economic sociology deal with the issues of social class in economic life, the role of taxation ('fiscal sociology') and imperialism (Schumpeter 1991). Schumpeter thought highly of these essays and they are all considered minor classics today.

But one can also find elements of economic sociology in some of Schumpeter's non-sociological writings. This goes for the famous analysis of entrepreneurship in *Theory of Economic Development,* not least the element of resistance from the environment that the entrepreneur usually confronts (Schumpeter 1934). Similarly in *Capitalism, Socialism and Democracy*, we find a sociological portrait of contemporary capitalism. The US economy was doing very well, according to Schumpeter, but its institutions were decaying (Schumpeter 1942).

Like Schumpeter, Karl Polanyi (1886–1964) came from the Austro-Hungarian Empire and ended his life on the American continent. Like Schumpeter, he wrote a famous book on capitalism – *The Great Transformation* – and contributed to the economic sociology of his days (Polanyi 1957). It is to Polanyi that we owe the term 'embeddedness', even if he used it in his own, very political sense: all economies had been embedded in politics and religion before the advent of capitalism, and were disembedded by the traumatic 'great transformation'. The political task of the day, in other words, was to re-embed the economy into political and human values.

Polanyi covered historical distances with great ease and was as much at home in ancient Babylonia as in 19th-century Britain or 20th-century United States. The scope of his knowledge about the economy is also reflected in one of his most useful sets of categories: the concepts of reciprocity, redistribution and exchange (for example, Polanyi 1971). In a kinship situation, for example, reciprocity may be used as a way of distributing resources. A political centre, like the state, would in contrast redistribute resources; and a market distributes resources through exchange. Most economic systems draw on each of these three ways of distributing resources, with their corporate sectors ('exchange'), state sectors ('redistribution') and household sectors ('reciprocity').

Talcott Parsons (1902–1979) had begun his career as an economist, only to switch to sociology, since he thought that utilitarian thought was unable to properly capture the structure of modern society. Parsons argued for a general systems perspective in social theory, and suggested in *Economy and Society* (together with Neil Smelser) that the economy should be conceptualized as a sub-system of the general system of society (Parsons and Smelser 1956). Just as each society has to have a distinct goal ('Polity') and a value-system ('Latent-Pattern-Maintenance'), it also has to adapt to nature and reality ('Economy'). While it is part of society, the economy is also its own society, with a 'polity', 'latent-pattern-maintenance', and so on.

New Economic Sociology

Around the mid-1980s American sociologists suddenly started to become interested in economic sociology, and it is this development that is generally known as 'new economic sociology'. One article in particular operated as a catalyst in this process, and that is Mark Granovetter's 'Economic action and social structure: the problem of embeddedness' (1985). Its central argument is that all economic actions are embedded in personal networks, and it is this quality that brings them into the sociologist's domain. While this message was important enough in itself, the article's implicit or subliminal message that sociology had neglected a whole area of social life which lent itself to sociological analysis, namely, the economy, also explains its great impact. Since sociological skills had not been applied to economic problems, sociologists might also be able to solve a number of important puzzles that the economists had failed to do, according to Granovetter.

Since the mid-1980s economic sociology has advanced steadily, and it is now fully institutionalized in the United States. It is routinely taught in sociology departments in all the major universities and also has a strong presence among the major journals of the profession. The American Sociological Association has a special section for economic sociology; a number of readers have been published as well as a huge handbook (Smelser and Swedberg 1994; 2005).

Economic sociology is becoming increasingly popular and accepted in Europe as well, though in

a somewhat different form than in the United States, which is only natural given the various national traditions in sociology. While interesting contributions can be found in many European countries, it is especially in France that one can find highly original contributions that stand up well to international competition (for England, see for example Dodd 1994; for Scotland, MacKenzie 2003; for Germany, Beckert 2004; for Italy, Trigilia 2002; and for Sweden, Aspers 2001).

The three key figures in French economic sociology are Pierre Bourdieu, Luc Boltanski and Michel Callon (see also the works of Lebaron 2000, and Steiner 2005). Bourdieu (1930–2002) has, among other things, analysed consumption in an innovative manner in his celebrated study *Distinction* (1986); he has also sketched a whole programme for economic sociology, drawing on his three key concepts of habitus, field, and different types of capitals (Bourdieu 1979; 2005). Luc Boltanski has contributed to the discussion of modern capitalism through an important study of class formation and also co-authored a provocative volume on 'the new spirit of capitalism' (Boltanski 1987; Boltanski and Chiapello 1999). And Michel Callon (1998) has introduced the so-called theory of performativity or the idea that economic theory may be as successful as an explanatory approach for the simple reason that it analyses phenomena that it has helped to create in the first place.

The number of studies in economic sociology (books and articles) amounts to several thousand by now, which makes it hard to summarize its achievements. One way to convey a sense of this literature, however, would be to discuss the methods that are being used to gather and analyse data as well as some of the most important topics. That economic sociology indeed has a distinctive profile that sets it off from mainstream economics emerges very clearly from a discussion of these two themes.

The data that is being used in economic sociology has often been put together by the analyst, and it is considerably less common than in mainstream economics to draw on official data of the type that is produced by government agencies. One example is historical studies in economic sociology, as illustrated by Bruce Carruther's *City of Capital* (1996). The focus in this work is the emergence of one of the world's first financial markets, and the author draws heavily on various primary and secondary sources. In particular, Carruthers succeeds in showing that early trade in shares often followed party lines; that is, sellers were reluctant to trade with political opponents.

Comparative studies are long-standing in economic sociology and have also been popular in new economic sociology. In one of these, *Forging Industrial Policy*, Frank Dobbin (1994) compares the ways in which the railroad industry developed in the 19th century in the United States, Britain and France. The author shows that industrial policy has largely mirrored the general political culture in its approach to solving problems in each of these three countries. In the United States, there has been scepticism towards the state and reliance on the corporations; in France, the state has been the central actor; and in Britain there has been an attempt to protect the individual firm from competition as well as from interventions from the state. Dobbin claims to have found that there is no one best way of doing things. Rather, people generalize from how they themselves do things and proclaim this to be the universally rational way to proceed.

Economic sociologists also draw on ethnography and participant observation, two methods that allow the researcher to handle huge amounts of empirical detail and to approach things from the perspective of the actors. Michael Burawoy (1979), for example, worked as a shop steward in order to better understand how workers interact and deal with the demands of their work (especially boredom); and Mitchel Abolafia (1996; 1998) passed an examination as a stockbroker in order to better understand what goes on in various stock and bond exchanges.

By far the most significant single method used by economic sociologists today, however, is that of networks. This is a very flexible tool, which allows for quantification and therefore goes well with a large number of research tasks. It has been used, for example, to analyse the links that exist between corporations by virtue of having the same

individual on their boards (so-called interlocks). Through the resultant system of communication, various ways of doing things may be diffused. The so-called poison pill (a measure against hostile takeovers) has, for example, been shown to diffuse quickly among corporations linked by common board members (Davis 1991). That links between corporations are not to be understood exclusively in terms of instrumental actions may be exemplified by the fact that, when a board member resigns or dies, he or she is only replaced in something like half of the cases (Palmer 1983).

Using networks is also a popular way in economic sociology to approach collaboration between corporations as well as the relationship between firms and their customers and suppliers (see, for example, Gulati and Gargiulo 1999). The area where it has been most successful, however, may well be the labour market; and here the classic study is Mark Granovetter's *Getting a Job* (Granovetter 1974). While one may have thought that the most important source of assistance for a person seeking a job is that's person's closest friends and family ('strong ties'), in fact it is his or her more casual contacts ('weak ties'), whose number depends on how many jobs a person has had. The reason for this 'strength of weak ties' is simply that, whereas one's 'strong ties' all share the same information, 'weak ties' can provide access to new and varied information, including information about job opportunities.

In European economic sociology an attempt has also been made to expand the notion of networks to include not only people and organizations in the category of actors but also objects (so-called actor-network-theory; see, for example, Law and Hassard 1999). That objects can be actors in the conventional sense of this term is no doubt wrong; the weaker claim that objects can be part of networks is, however, more interesting. One may, for example, see a machine as a link between people, some objects may be used for communication between people, and so on – and all this can affect the structure of the network. More generally, the advocates of actor-network-theory also argue that the traditional approach of economists and sociologists tends totally to ignore the role that objects play in the economy and to focus exclusively on actions, social relations and the like. The perspective that argues for including objects in the analysis is usually referred to as 'materiality'.

When it comes to the topics that are often analysed, new economic sociologists have first and foremost tried to focus on economic institutions as opposed to phenomena situated at the boundary of, say, religion and the economy or politics and the economy. The reason for this has been a desire to take on truly 'economic' topics and go beyond the old division of labour between economics and sociology, when the former dealt with the economy and the latter with society minus the economy. As examples of this is the interest among contemporary economic sociologists in markets and corporations, which have attracted a large number of studies.

One type of study has attempted to develop a general model for markets that differs sharply from the standard economic model of the perfect market. The most prominent example of this is the work of Harrison White (1981; 2002) on so-called production markets, by which he roughly means industrial markets. Production markets, it is argued, differ from so-called exchange markets primarily because their participants have permanent roles as either sellers or buyers and do not switch between these two roles as is common in financial markets.

According to White, the typical production market holds about a dozen actors who closely follow what the other actors are up to. Markets come into being, White argues, precisely because economic actors position themselves in relation to the products of other actors. Prices are not set through demand and supply but by producers relating the revenue of their goods to the volume that is being sold. Individual markets, finally, are connected to each other in giant networks, either 'upstreams' (suppliers) or 'downstreams' (customers).

A number of studies of financial markets have also been carried out, and here the work of Donald MacKenzie is outstanding (for example, MacKenzie 2003; MacKenzie and Millo 2003). MacKenzie has picked up from Callon the theme of performativity, and he uses it, for example, in his

analysis of trade in options. The pricing of options was very difficult, the argument goes, until Black, Scholes and Merton suggested a solution for which the latter two would win the Nobel Prize in 1997. While this formula covers most cases with much precision, according to MacKenzie it does not cover all – and this was to have important consequences. Since this fact was not well understood, however, and since economic reality was mistaken for how it was portrayed in finance theory (performativity), there have been cases in which people were unprepared for what was happening (as in the case of Long-Term Capital Management). MacKenzie traces this development and also shows how actors have tried to protect themselves against exceptional cases by keeping a margin against the price predicted according to the Black–Scholes–Merton formula.

Economic sociologists have suggested several new ways to approach consumer markets. Viviana Zelizer (1979), for example, has analysed the growth of the market in life insurance in the United States and shown how the idea of putting a price on a human life initially attracted hostility, for religious reasons. But as people moved into the cities and religion had to adjust to new circumstances, a different view of life insurance emerged. Zelizer has recently also started to look at consumption among children, both how children are socialized into becoming consumers and the ways in which they themselves relate to objects and goods in their environment (Zelizer 2005).

DiMaggio and Louch (1998) have attempted to use networks to analyse consumption. While it is well known that people will turn to others in their surroundings to find out where to buy something, and which merchants, traders and so on are reliable ('search embeddedness'), DiMaggio and Louch examine situations in which people approach someone in their personal network in order to buy something ('within-networks exchange'). As it happens, this is quite common, especially infrequent purchases of the type that involve legal services, home repair maintenance and the buying of a car or a home.

The number of studies in economic sociology that deal with corporations is very great, but a few studies nonetheless stand out. One of these is Mark Granovetter's pioneering 1994 article on business groups. Against R.H. Coase, Granovetter argues that it is not so much the existence of the individual firm that needs to be explained but the common phenomenon of groups of firms. In many countries, such as India, South Korea and Japan, these business groups control large parts of the economy, but have not received the scholarly attention that they deserve. The impact of business groups in the United States is not clear from Granovetter's work, except that US antitrust legislation has ruled out some common forms of this phenomenon.

The business groups that Granovetter studies lend themselves to a networks approach, and so do the corporations that Ronald Burt (1983) has analysed in his study of US industrial markets. Each firm, according to Burt, can be conceptualized as situated at the centre of a network in which there are a number of competitors, suppliers and customers. The fewer competitors there are, the more suppliers, and the more customers, the more the corporation is characterized by 'structural independence'. And with more structural independence comes more profit, as Burt shows.

The emphasis on corporations in interaction, as opposed to the single corporation, is also obvious in another landmark study in economic sociology, *Regional Advantage* by AnnaLee Saxenian (1994). Following Alfred Marshall in analysing industrial districts, Saxenian carries out a comparative study of the computer industry during the post-war period in Silicon Valley and the area around Route 128 in Boston. Silicon Valley has clearly overtaken Route 128 during recent decades, and the reason for this, according to Saxenian, has to do with the nature of the interaction in the two regions. While in Route 128 the corporations are loath to cooperate, rely on banks for finance, and prosecute employees who switch to competitors, in Silicon Valley there is plenty of cooperation, finance comes from venture capital firms, and employees are free to switch as they like. A much more decentralized and flexible form of entrepreneurship, in brief, has emerged in Silicon Valley.

Saxenian's fascination with entrepreneurship is shared by many economic sociologists. While

she argues that a radical decentralized industrial region represents the best conditions for entrepreneurship, there exist other perspectives as well. Granovetter, for example, argues that entrepreneurs often come from those parts of the social system which are far away from the controlling centre (for example, Granovetter 2005). While this may be termed a theory of peripheral entrepreneurship, Granovetter suggests several other situations that are favorable to entrepreneurship. An entrepreneur may, for example, be someone who crosses a social boundary in society and thereby becomes the first to unite resources from two otherwise separated regions (for example, Granovetter 1995). On immigration, Granovetter also points out that some ethnic groups that are not entrepreneurial in their country of origin may be highly entrepreneurial in their new country because they often leave parts of the extended family behind (Granovetter 1995). This means that they do not have to provide jobs for their relations or share their wealth with relatives.

Economic sociologists have been very active in studying ethnic entrepreneurship, (for example, Light 2005). Ethnic entrepreneurs, for example, often have to overcome the fact that their initial market consists of their countrymen ('the ethnic market'), and that they will have to go beyond this market if they are to expand. In many cases they have become entrepreneurs simply because they have no other way of making a living ('forced entrepreneurship').

Economic sociologists have also emphasized the collective nature of entrepreneurship and attempted to explode the myth of the creative Schumpeterian individual. One important example of this can be found in the research by Rosabeth Moss Kanter (1983) on entrepreneurship within the corporation, so-called intrapreneurship. Through a combination of ethnographic studies and survey research, Kanter has attempted to show the conditions under which it is possible to put together creative and entrepreneurial groups in modern corporations. Someone has to suggest the creation of such groups and provide them with resources and legitimacy. The group also has to be defended from outside intervention while it operates, internal conflicts have to be solved, and so on. According to Kanter, this type of group is common among modern corporations.

While economic sociologists have been unable to present a general theory of entrepreneurship, it is nonetheless clear that a number of insights have been accumulated. Economic sociologists are also expanding their work into such topics as social entrepreneurship and the diffusion of courses among business schools (for example, Swedberg 2000).

Concluding Remarks

Economic sociology is currently in a very active phase of its development, and all signs indicate that this trend will continue. Economic sociologists are also gradually expanding their range of topics of study. There has recently, for example, been an attempt to introduce law into the analysis, and some economic sociologists are trying to formulate a position on the relationship between the economy and technology. Some economic sociologists are also in the process of investigating the role of emotions in the economy; and there is a growing number of studies of gender and the economy. What all of this adds up to, again, is a steady growth of studies in economic sociology and a confirmation that economic sociology is established as a distinct and accepted area of sociology. But it remains to be seen whether economic sociology will be able to make inroads into economics itself and gain respect from economists, along the lines of, say, behavioural economics.

See Also

- ▶ Akerlof, George Arthur (Born 1940)
- ▶ Cartels
- ▶ Entrepreneurship

Bibliography

Abolafia, M. 1996. *Making markets: Opportunism and restraint on Wall Street*. Cambridge, MA: Harvard University Press.

Abolafia, M. 1998. Markets as culture: An ethnographic approach. In *The laws of the markets*, ed. M. Callon. Oxford: Blackwell.

Aspers, P. 2001. *A market in vogue: A study of fashion photography in Sweden*. Stockholm: City University Press.

Beckert, J. 2004. *Unverdients Vermögen. Soziologie des Erbrechtes*. Frankfurt: Campus Verlag.

Boltanski, L. 1987. *The making of a class: Cadres in French society.* Cambridge: Cambridge University Press.

Boltanski, L., and E. Chiapello. 1999. *Le Nouvel Esprit du Capitalisme*. Paris: Gallimard.

Bourdieu, P. 1979. *Algeria 1960*. Cambridge: Cambridge University Press.

Bourdieu, P. 1986. *Distinction: A social critique of the judgment of taste*. London: Routledge.

Bourdieu, P. 2005. Principles of an economic anthropology. In *The handbook of economic sociology*, 2nd ed, ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.

Burawoy, M. 1979. *Manufacturing consent: Changes in the labor process under monopoly capitalism*. Chicago: University of Chicago Press.

Burt, R. 1983. *Corporate profits and cooptation: Networks of market constraints and directorate ties in the American economy.* New York: Academic.

Burt, R. 1992. *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Callon, M. (ed.). 1998. *The laws of the markets*. Oxford: Blackwell.

Carruthers, B. 1996. *City of capital: Politics and markets in the English financial revolution*. Princeton: Princeton University Press.

Davis, G. 1991. Agents without principles? The spread of the poison pill throughout the intercorporate network. *Administrative Science Quarterly* 36: 583–613.

DiMaggio, P., and H. Louch. 1998. Socially embedded consumer transactions: For what kind of purchases do people most often use networks? *American Sociological Review* 63: 619–637.

Dobbin, F. 1994. *Forging industrial policy: The United States, Britain, and France in the railway age*. New York: Cambridge University Press.

Dodd, N. 1994. *The sociology of money: Economics, reason and contemporary society.* Cambridge: Polity Press.

Durkheim, E. 1893. *The division of labor in society*, 1984. New York: Free Press.

Granovetter, M. 1974. *Getting a job: A study of contacts and careers*. Cambridge, MA: Harvard University Press.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91: 481–510.

Granovetter, M. 1994. Business groups. In *The Handbook of economic sociology*, ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.

Granovetter, M. 1995. The economic sociology of firms and entrepreneurship. In *The economic sociology of immigration*, ed. A. Portes. New York: Russell Sage Foundation.

Granovetter, M. 2005. The impact of social structure on economic outcomes. *Journal of Economic Perspectives* 19(1): 33–50.

Gulati, R., and M. Gargiulo. 1999. Where do interorganizational networks come from? *American Journal of Sociology* 104: 1439–1493.

Jevons, S. 1879. *The theory of political economy*, 5th ed. New York: Augustus M. Kelley, 1965.

Kanter, R. 1983. *The change masters: Innovation and entrepreneurship in America*. New York: Simon and Schuster.

Law, J., and J. Hassard (eds.). 1999. *Actor network theory and after.* Oxford: Blackwell.

Lebaron, F. 2000. *La Croyance Economique: Les Economistes entre Science et Politique*. Paris: Seuil.

Light, I. 2005. The ethnic economy. In *The handbook of economic sociology*, 2nd ed, ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.

MacKenzie, D. 2003. Long-term capital management and the sociology of arbitrage. *Economy and Society* 32: 349–380.

MacKenzie, D., and Y. Millo. 2003. Constructing a market, performing theory: The historical sociology of a financial derivatives exchange. *American Journal of Sociology* 109: 107–145.

Marshall, G. 1982. *In search of the spirit of capitalism: An essay on Max Weber's protestant ethic thesis*. London: Hutchinson.

Palmer, D. 1983. Broken ties: Interlocking directorates and intercorporate coordination. *Administrative Science Quarterly* 28: 40–55.

Parsons, T., and N. Smelser. 1956. *Economy and society: A study in the integration of economic and social theory.* New York: The Free Press.

Polanyi, K. 1957. *The great transformation*. Boston: Beacon Hill.

Polanyi, K. 1971. The economy as instituted process. In *Trade and market in the early empires*, ed. K. Polanyi, C. Arensberg, and H. Pearson. Chicago: Henry Regnery.

Saxenian, A. 1994. *Regional advantage: Culture and competition in silicon valley and route 128*. Cambridge, MA: Harvard University Press.

Schumpeter, J. 1934. *The theory of economic development*. Cambridge, MA: Harvard University Press.

Schumpeter, J. 1954. *History of economic analysis*. London: Allen & Unwin.

Schumpeter, J. 1991. In *The economics and sociology of capitalism*, ed. R. Swedberg. Princeton: Princeton University Press.

Schumpeter, J. 1942. *Capitalism, socialism and democracy.* London: Routledge, 1994.

Simmel, G. 1900. *The philosophy of money.* London: Routledge, 1978.

Simmel, G. 1908. Competition. In *Conflict and the web of group-affiliation*. New York: The Free Press, 1955.
Smelser, N., and R. Swedberg (eds.). 1994. *The handbook of economic sociology*. Princeton: Princeton University Press.
Smelser, N., and R. Swedberg (eds.). 2005. *The handbook of economic sociology*, 2nd ed. Princeton: Princeton University Press.
Steiner, P. 2005. *L'Ecole Durkheimienne et l'Economie*. Geneva: Droz.
Swedberg, R. 1998. *Max Weber and the idea of economic sociology*. Princeton: Princeton University Press.
Swedberg, R. 2000. The social science view of entrepreneurship. In *Entrepreneurship: The social science view*, ed. R. Swedberg. Oxford: Oxford University Press.
Trigilia, C. 2002. *Economic sociology: State, market and society in modern capitalism*. Oxford: Blackwell.
Weber, M. 1895. The national state and economic policy (Freiburg Address). *Economy and Society* 9: 428–449 (1980).
Weber, M. 1905. *The protestant ethic and the spirit of capitalism*. New York: Charles Scribner's Sons, 1958.
Weber, M. 1909. *The Agrarian sociology of ancient civilizations*. London: New Left Books, 1976.
Weber, M. 1914. *Economy and society: An outline of interpretive sociology*. Berkeley: University of California Press, 1978.
White, H. 1981. Where do markets come from? *American Journal of Sociology* 87: 517–547.
White, H. 2002. *Markets from networks: Socioeconomic models of production*. Princeton: Princeton University Press.
Whyte, W. 1955. *Money and motivation*. New York: Harper.
Zelizer, V. 1979. *Morals and markets: The development of life insurance in the United States*. New York: Columbia University Press.
Zelizer, V. 2005. Culture and consumption. In *The handbook of economic sociology*, 2nd ed, ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.

Economic Surplus and the Equimarginal Principle

Maurice Allais

Keywords

Allais, M.; Economic surplus; Economy of markets; Equimarginal principle; General equilibrium; Marginalism; Marginal equivalence; Marginal rate of substitution; Maximum efficiency; Preference index; Production functions; Surplus

JEL Classifications

D5

Marginal analysis is actually only a particular case of a more general theory, the theory of surpluses and the economy of markets, which, if considered first, facilitates the discussion of the equimarginal principle.

The General Theory of Surpluses and the Economy of Markets: Fundamental Concepts and Theorems

To simplify the exposition, it is assumed that one good (U), enters all preference and production functions, and that its quantity can vary continuously. Except for the hypothesis of continuity with respect to this good (U), the discussion in this first part is free of any restrictive hypothesis of continuity, differentiability or convexity for the goods (V),. . .,(W) considered, and the preference indexes and production functions. (For an exposition of the following theory in the case where no one good plays a particular role, see Allais 1985, Section II, pp. 139–41.)

Structural Conditions

The needs of every unit of consumption, individual or collective, can be entirely defined by considering a preference index

$$I_i = f_i(U_i, V_i, \ldots, W_i) \tag{1}$$

increasing as it passes from a given situation to one it finds preferable. Every quantity V_i is counted positively if it refers to a consumption, negatively if it refers to a service supplied.

The set of feasible techniques for a unit of production j can be represented by a condition of the form

$$f_j(U_j, V_j, \ldots, W_j) \geq 0$$

where every quantity V_j is considered as representing a consumption or an output depending on whether it is positive or negative.

The extreme points corresponding to the boundary between possible and impossible situations represent states of maximum efficiency for the production unit considered. They may be represented by the condition

$$f_j(U_j, V_j, \ldots, W_j) = 0. \tag{2}$$

The function f_j may be called the production function. It is defined up to any transformation which leaves its sign unchanged.

From a technical point of view, maximum efficiency implies quite specific conditions. If, for instance, one considers a production technique $A = A\ (X, Y,..., Z)$ and if n production units are technically preferable to a single one, we should have (Allais 1943, pp. 187–8; 1981, pp. 319–22)

$$\sum_j A(X_j, Y_j, \ldots, Z_j) > A\left[\sum_j X_j, \sum_j Y_j, \ldots, \sum_j Z_j\right]. \tag{3}$$

In the opposite case we have

$$A\left[\sum_j X_j, \sum_j Y_j, \ldots, \sum_j Z_j\right] > \sum_j A(X_j, Y_j, \ldots, Z_j). \tag{3*}$$

An industry is referred to as differentiated if the use of distinct production units is technically more advantageous than the concentration of all production operations into a single production unit. It is called non-differentiated in the opposite case. Conditions (3) and (3*) are two particular illustrations of differentiation (Allais 1943, p. 637).

From inequality (3) it is possible to show that the whole production function of a differentiated industry is asymptotically homogeneous. In this case ($n \gg 1$) there is quasihomogeneity (Allais 1943, pp. 201–6; 1974b).

Distributable Surplus Corresponding to a Given Modification of the Economy

The distributable surplus σ_u relative to a good (U) and to a realizable modification of the economy which leaves all preference indexes unchanged is defined as the quantity of that good which can be released following this shift (Allais 1943, pp. 610–16). The surplus considered here differs essentially from the concepts of consumer surplus as normally considered in the literature (for example, Samuelson 1947, pp. 195–202; Blaug 1985, pp. 355–70; Allais 1981, pp. 297–8, and 1985, nn. 12–13).

Let us consider an initial state $(\mathscr{E}_1)$ characterized by consumption values U_i, V_I,...,W_i and U_j, V_j,..., W_j (positive or negative) of the different units of consumption and production. We have

$$\Sigma_i U_i + \Sigma_j U_j = U_0;$$
$$\Sigma_i V_i + \Sigma_j V_j = V_0; \ldots; \Sigma_i W_i + \Sigma_j W_j = W_0 \tag{4}$$

where U_0, V_0,..., W_0 designate available resources. Let $(\delta\mathscr{E}_1)$ be a feasible modification of $(\mathscr{E}_1)$ characterized by finite variations δU_i, δV_i,..., δW_i, δU_j, δV_j,..., δW_j, and let

$$(\mathscr{E}_2) = (\mathscr{E}_1) + \delta(\mathscr{E}_1)$$

represent the new state.

According to (4) we naturally have

$$\sum_i \delta V_i + \sum_j \delta V_j = 0$$

for every good (U), (V),..., (W). From (2) we also have for every unit of production j

$$f_j(U_j + \delta U_j, V_j + \delta V_j, \ldots, W_j + \delta W_j) = 0.$$

According to (1) the preference indexes become

$$I_i + \delta I_i = f_i(U_i + \delta U_i, V_i + \delta V_i, \ldots, W_i + \delta W_i).$$

The δI_i can be positive, zero, or negative.

Let us now define a third state $(\mathscr{E}_3)$ by the condition that by the modification $-\delta\sigma_{ui}$ of just the quantities $U_i + \delta U_i$ all the preference indexes return to their initial values.

We then have the conditions

$$f_i(U_i + \delta U_i - \delta\sigma_{ui}, V_i + \delta V_i, \ldots, W_i + \delta W_i) = f_i(U_i, V_i, \ldots, W_i). \tag{5}$$

The state $(\mathscr{E}_3)$ can be termed 'isohedonous' with the state $(\mathscr{E}_1)$. In passing from $(\mathscr{E}_1)$ to $(\mathscr{E}_3)$ the quantity

$$\delta\sigma_u = \sum_i \delta\sigma_{ui} \tag{6}$$

of the good (U) is released, as all the units of consumption find themselves again in situations which they consider equivalent, since their preference indexes return to the same values (Allais 1943, pp. 637–8).

The surplus $\delta\sigma_u$ has been released during the passage from $(\mathscr{E}_1)$ to $(\mathscr{E}_3)$. It may then be considered that in the situation $(\mathscr{E}_1)$ this surplus was both realizable and distributable. It may further be considered that in passing from $(\mathscr{E}_1)$ to $(\mathscr{E}_2)$, it has in effect been distributed.

The distributable surplus thus defined covers the whole economy, but this definition can be used for any group of agents. It is necessary only to consider the functions f_i and f_j and the resources relating to this group in the preceding relations.

Any exchange system, with the corresponding production operations it implies, is deemed 'advantageous' when a distributable surplus is achieved and distributed, so that the preference index of any consumption unit concerned increases. If an exchange and production system is advantageous, there must be at least one system of prices which allows it, the prices used by each pair of agents being specific to them. The distribution of the realized surplus between agents is determined by the system of prices used in the exchanges between them.

Conditions of Equilibrium and Maximum Efficiency

In essence all economic operations of whatever type may be considered as reducing to the search for, the achievement of, and the distribution of surpluses. Thus stable general economic equilibrium exists if, and only if, in the situation under consideration, there is no realizable surplus, which means

$$\delta\sigma_u \leq 0 \tag{7}$$

for all feasible modifications of the economy (Allais 1943, pp. 606–12).

In such a situation the distributable surplus is zero or negative for all possible modifications of the economy compatible with its structural relations, and it is impossible to find any set of prices that would permit effective bilateral or multilateral exchanges (accompanied by the implied production operations) which are advantageous to all the agents concerned.

A situation of maximum efficiency can be defined as a situation in which it is impossible to improve the situation of some people without undermining that of others, i.e. to increase certain preference indices without decreasing others. The set of states of maximum efficiency represents the boundary between the possible and the impossible (Fig. 1).

From those definitions of the situations of maximum efficiency and stable general economic equilibrium, it follows, with the greatest generality and without any restrictive hypothesis of continuity, differentiability or convexity, except for the common good (U), that:

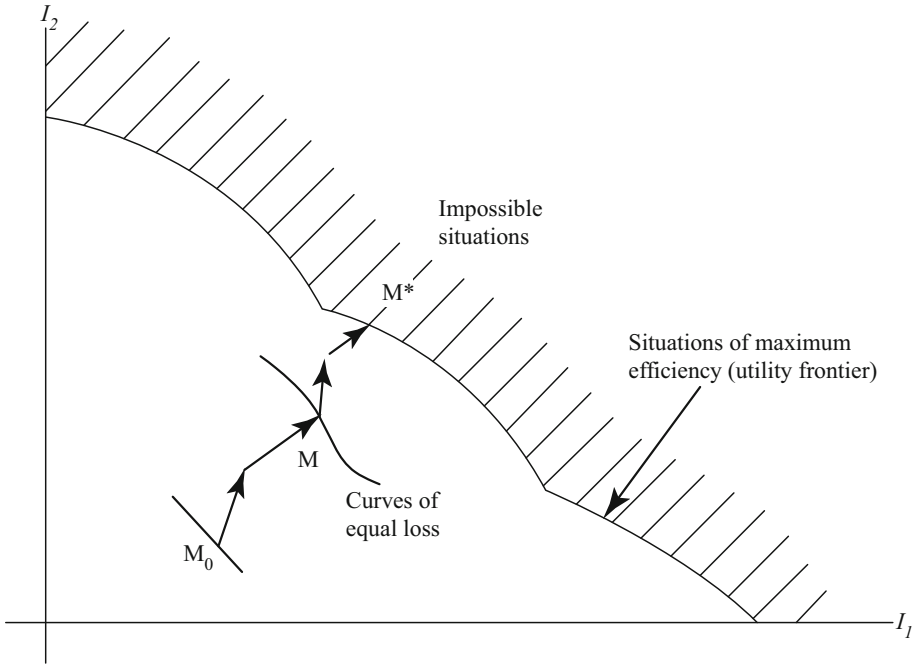
> Any state of stable general economic equilibrium is one of maximum efficiency (*First theorem of equivalence*). Any state of maximum efficiency is one of stable general economic equilibrium (*Second theorem of equivalence*).

Since there can be no stable general economic equilibrium if there is any distributable surplus, every state of stable general economic equilibrium is a state of maximum efficiency. Conversely, if there is maximum efficiency, there is no realizable surplus which could be used to increase at least one preference index without decreasing the others, and consequently, every state of maximum efficiency is a state of stable general economic equilibrium.

Because of the theorems of equivalence, the terms 'conditions of stable general economic equilibrium' and 'conditions of maximum efficiency' are used interchangeably below.

The Dynamic Process of the Economy: Decentralized Search for Surpluses

In their essence all economic operations, whatever they may be, can be thought of as boiling down to the pursuit, realization and allocation of distributable surpluses. The corresponding



Economic Surplus and the Equimarginal Principle, Fig. 1 Process of dynamic evolution. Illustrative diagram

model is the Allais model of the economy of markets (1967), defined by the fundamental rule that every agent tries to find one or several other agents ready to accept at specific prices a bilateral or multilateral exchange (accompanied by corresponding production decisions) which will release a positive surplus that can be shared out, and which is realized and distributed once discovered. Thus the evolution of the market's economy is characterized by the condition

$$\delta I_i \geq 0$$

for every consumption unit.

Since in the evolution of an economy of markets surpluses are constantly being realized and allocated, the preference indexes of the consumption units are never decreasing, at the same time as some are increasing. This means that for a given structure, that is to say, for given preferences, resources, and technical know-how, the working of an economy of markets tends to bring it nearer and nearer to a state of stable general economic equilibrium, hence a state of maximum efficiency (Fig. 1), which is the third fundamental theorem.

Naturally such evolution takes place only if sufficient information exists about the actual possibilities of realizing surpluses.

To any given initial situation whatsoever, assumed not to be a situation of equilibrium, there corresponds an infinite number of possible equilibrium situations, each corresponding to a particular path and each satisfying the general condition that no index of preference should take on a lower value than in the initial situation (Fig. 1).

Economic Loss

The loss σ_u^* which is associated with a given situation is defined as the greatest quantity of the good (U) which can be released in a transformation of the economy for which all the preference indexes remain unchanged (Fig. 1) (Allais 1943, pp. 638–49).

It is a well determined function

$$\sigma_u^* = F[I_1, I_2, \ldots, I_n, U_0, V_0, \ldots, W_0] \quad (8)$$

of the preference indexes I_i and of the resources V_0 which characterize this situation. The loss σ_u^* is an

indicator of inefficiency, and $-\sigma_u^*$ an indicator of the efficiency of the economy as a whole.

The loss is minimum and nil in every state of maximum efficiency, and positive in every feasible situation which is not a state of maximum efficiency. It decreases in any modification of the economy, whereby some preference indexes increase, others remaining unchanged, or whereby some surpluses are released with no decline in some preference indexes.

Paths to States of Economic Equilibrium and Maximum Efficiency

Since the preference indices I_i are continuous functions of the quantities U_i of the common good (U), the boundary between the possible and the impossible situations in the hyperspace of preference indexes is constituted by a continuous surface. On this surface the loss σ_u^* is nil. This representation allows an immediate demonstration by simple topological considerations of propositions whose proof would otherwise be very difficult. (The paternity of this representation has been unduly attributed to P. Samuelson 1950, but it was in fact published for the first time in Allais 1943, and systematically used by Allais in later years especially 1945 and 1947; see Allais 1971, n.11, p. 385; and 1974a, n.18, pp. 176–7.)

For every feasible situation which is not a state of maximum efficiency, represented by a point such as M_0, there are an infinity of realizable displacements M_0M enabling a situation of maximum efficiency M^* to be approached, such that all the preference indexes have greater values than in the initial situation M_0.

Figure 1 presents an illustration of the process of dynamic evolution by releasing and sharing out of surpluses during which the loss σ_u^* is constantly decreasing (Allais 1943, 1974b, and 1981, p. 121).

The Changing Structure of the Economy

As psychological patterns vary, as techniques are improved, or as new resources are discovered (or existing resources depleted), the set of situations of maximum efficiency relative to the indexes of preference constantly undergoes change over time. Consequently, situations of equilibrium and maximum efficiency are never reached, and what is really important is to determine the rules of the game which must be applied to come constantly closer to them as rapidly as possible. At a given time t, if information is sufficient and if the adjustments are sufficiently rapid, the point representing the economy will never be very far from the maximum efficiency surface of that time t.

General Comment

An economy of markets can be defined as one in which the agents – consumption, production, and arbitrage units – coexist and are free to undertake any exchange transaction or production operation which can result in rendering some distributable surplus available. The principle of the market economy is that any surplus realized is shared among the operators involved. How the surpluses achieved are shared out depends on the specific systems of prices used in the exchanges between the agents concerned. The prices used are always specific to the exchange and production operations considered and there is never a unique system of prices used in common by all the agents.

Diagrammatic representation like that of Fig. 1 reveals clearly three basic facts:

1. There is an infinity of situations of maximum efficiency corresponding to a given initial situation characterized by some distribution of property.
2. To each situation of maximum efficiency there corresponds a final distribution of property.
3. This final distribution depends on the initial situation and the distribution of surpluses in the course of the transition.

Thus there is a very strong interdependence between the point of view of efficiency corresponding to the discovery and realization of surpluses and the ethical point of view corresponding to their sharing.

In any event, since only what is produced can be shared, the incentive stemming from the partial or total appropriation of the surpluses by the various agents appears as a fundamental factor for the functioning of the economy of markets.

On the general theory of surpluses and the economy of markets in the general case, and on the fundamental theorems see Allais (1943, pp. 112–77; 181–211; 604–56; 1967, § 8–65; 1968a, vol. 2; 1968b, 1971, 1974a, 1981, pp. 27–48; 1985).

The Equimarginal Principle

Continuity and Differentiability

The preceding definitions and theorems are very general and do not make any hypothesis of continuity, derivability or convexity, except the hypothesis of continuity for the common good (U).

We now assume in addition only that all the quantities and functions considered are continuous and that all functions have first and second order derivatives, the following developments being totally independent of any hypothesis of general convexity.

From the sign conventions adopted earlier it follows that for any *i, j* and *V*

$$f'_{iv} = \partial f_i/\partial V_i \geq 0,\ f'_{jv} = \partial f_j/\partial V_j \geq 0.$$

The second partial derivatives are written

$$f''_{ivw} = \partial^2 f_i/\partial V_i \delta W_i, \quad f''_{jvw} = \partial^2 f_j/\partial V_j \delta W_j\,.$$

In the following, the symbol $\overline{d^2 g}$ represents the second differential

$$\overline{d^2 g} = \sum_U^W g''_{v^2} \mathrm{d}V^2 + 2 \sum_{U,V} g''_{vw} \mathrm{d}V\,\mathrm{d}W$$

of a function $g(U, V, \ldots, W)$ when all parameters in that function are taken as independent, while the symbol $\overline{d^2 g}_u$ represents what this second differential becomes after du has been replaced by its expression derived from

$$\mathrm{d}g = \sum_U^W g'_v \mathrm{d}V = 0$$

(Allais 1968a, vol. 2, pp. 77–8; 1973b, pp. 151–5; 1981, pp. 688–9).

Convexity and Concavity

The local properties of diminishing or increasing marginal returns are related to local conditions of convexity or concavity. Convexity is defined as follows:

Ordinal fields of preference A field of choice is said to be convex in the whole space (postulate of general convexity) if, at all points of the field, the condition

$$I(M_0) \leq I(M_1)$$

entails

$$I(M_0) \leq I(M)$$

with

$$M = \lambda M_0 + (1 - \lambda) M_1 0 < \lambda < 1.$$

There is local convexity at M_0 if this condition is satisfied only for

$$|\mathrm{M_0 M_1}| < \varepsilon$$

where ε is a given positive number.

When differentiability is assumed local convexity implies.

$$\overline{d^2 f_{iu}} \leq 0 \quad \text{for} \quad \mathrm{d}f_1 = 0\,.$$

Fields of production A field of production is said to be convex over the whole space (postulate of general convexity) if, for any two possible points M_0 and M_1, the centre of gravity defined by the relation

$$M = \lambda M_0 + (1 - \lambda) M_1$$

is likewise a possible point for.

$$0 < \lambda < 1.$$

Local convexity obtains at M_0 if the preceding condition is satisfied only for

$$|M_0 M_1| < \varepsilon$$

where ε is a given positive number.

When differentiability is assumed, local convexity implies

$$\overline{d^2 f_{iu}} \leq 0 \quad \text{for} \quad \mathrm{d}f_1 = 0 \cdot$$

In fact there is no production operation that does not begin by providing increasing marginal returns, and it is only beyond a certain threshold that diminishing marginal returns are observed. That is a general physical law of nature (Allais 1943, pp. 193–5; 1968a, vol. 2, pp. 68–96; 1971, pp. 362–4; 1974a, pp. 153–7). Similarly it can be considered as an introspective datum that psychological returns begin by increasing but in the end always decrease beyond certain threshold values. That is a general psychological law (Allais 1968a, vol. 2, pp. 109–38; 1971, pp. 360–2; 1974a, pp. 153–5). These are two fundamental properties of fields of choice and production. They rule out the postulate of general convexity which is generally accepted in the contemporary literature.

Generation of Distributable Surplus

Consider any economic state $(\mathscr{E})$ and a realizable modification $(\delta\mathscr{E})$ such that all the preference indexes I_i remain constant (isohedonous modification). Let the conditions of constancy of these indexes and the conditions corresponding to the production functions be written in the same general form

$$g_k(U_k, V_k, \ldots, W_k) = 0 \tag{9}$$

where U_k, V_k,..., W_k represent the consumption of both consumption and production units. By convention, any quantity V_k, if positive, represents consumptions, either by a consumption or a production unit. For any production or consumption unit, any parameter V_k, if negative, represents production of a good or a service.

Let $\mathrm{d}U_k$, $\mathrm{d}V_k$,..., $\mathrm{d}W_k$, be the first order differentials of the variations δU_k, δV_k,..., δW_k of consumptions U_k, V_k,..., W_k in the displacement $(\delta\mathscr{E})$. From (9), we have

$$g'_{ku}\mathrm{d}U_k + g'_{kv}\mathrm{d}V_k + \cdots + g'_{kw}\mathrm{d}W_k = 0. \tag{10}$$

Let δV_{kl} be the quantity of (V) received by the consumption or the production unit k from the consumption or production unit l. By definition, we have

$$\delta V_k = \sum_{k \neq 1} \delta V_{kl} \tag{11}$$

$$\delta V_{lk} = -\delta V_{kl}. \tag{12}$$

Assuming that the displacement $(\delta\mathscr{E})$ is such that

$$\sum_k \delta V_k = 0, \ldots, \sum_k \delta W_k = 0. \tag{13}$$

Let

$$\varepsilon^k_{v,u} = g'_{kv}/g'_{ku}. \tag{14}$$

The ratio $\varepsilon^k_{v,u}$ is the coefficient of marginal equivalence (or marginal rate of substitution) of goods (V) and (U) for agent k (Allais 1943, pp. 609–10, and 617–21).

From (10) and (14) we have the relation (15)

$$\mathrm{d}U_k = -\left[E^k_{vu}\mathrm{d}V_k + \cdots + E^k_{wu}\mathrm{d}W_k\right] \tag{15}$$

between the first order differential $\mathrm{d}U_k$, $\mathrm{d}V_k$,..., $\mathrm{d}W_k$.

If $\mathrm{d}U_k$ is positive, agent k receives a quantity $\mathrm{d}U_k$ to within the second order. If $\mathrm{d}U_k$ is negative, agent k supplies a quantity $-\mathrm{d}U_k$ to within the second order.

From the condition (13), it follows that the displacement considered releases a global distributable surplus

$$\delta\sigma_u = -\sum_k \delta U_k$$

representing the excess of the quantities supplied over the quantities received of good (U) whose first order differential is

$$\delta\sigma_u = -\sum_k \mathrm{d}U_k.$$

From (11) and (15)

$$\mathrm{d}U_k = -\sum_{v}^{w}\left[\varepsilon_{vu}^{k}\sum_{\substack{k,l\\k<l}}\mathrm{d}V_{kl}\right]$$

and from (12), we have (Allais 1952c, p. 31; 1968a, vol. 2, p. 174; 1981, p. 88)

$$\mathrm{d}\sigma_u = \sum_{v}^{w}\sum_{\substack{k,l\\k<l}}\left(\varepsilon_{vu}^{k} - \varepsilon_{kl}^{l}\right)\mathrm{d}V_{kl}. \tag{16}$$

According to definitions (5) and (6) $\mathrm{d}\sigma_u$ is the first differential of the global distributable surplus $\delta\sigma_u$ released in the displacement considered. For all economic agents the unit of value is defined by condition $u_k = u = 1$. The marginal values v_k,..., w_k of goods (V),. . ., (W) for unit k are defined with respect to the u_k by the relations

$$\frac{g'_{ku}}{u_k} = \frac{g'_{kv}}{v_k} = \cdots = \frac{g'_{kw}}{w_k} \tag{17}$$

$$u_k = u = 1. \tag{18}$$

Under the adopted sign convention, all the v_k are positive. We have from (14) and (18)

$$\varepsilon_{vu}^{k} = v_k \tag{19}$$

and relation (16) is written

$$\mathrm{d}\sigma_u = \sum_{v}^{w}\sum_{\substack{k,l\\k<l}}(v_k - v_l)\mathrm{d}V_{kl} \tag{20}$$

where v_k and v_l are the marginal values of good (V) for units k and l. This summation covers all agents, both consumption and production units. It can thus be seen that all the differences between the marginal values in the situation $\mathscr{E}$ can give rise to the release of potential surpluses which can be released and distributed.

The meaning of relation (20) is immediate. Thus if $v_k > v_l$ the relative value of good (V) is higher for agent k than for agent l. The transfer of a positive quantity $\mathrm{d}V_{kl}$ of good (V) from agent l to agent k therefore creates an additional positive value

$$\mathrm{d}\sigma_{ukl} = (v_k - v_l)\mathrm{d}V_{kl}.$$

If in this 'isohedone' transformation surpluses are released, all positive, they can be distributed in such a way as to increase all preference indexes. In such a modification of the economy, the maximum distributable surplus diminishes, and the point representing the economic situation considered moves closer to the surface of maximum efficiency in the hyperspace of preference indexes. Naturally, for this condition to obtain, the corresponding exchanges and the changes of the consumptions and productions they imply in the production system, must effectively occur.

Psychological Values and Marginal Psychological Values

Naturally, the v_k are only marginal values for the agents. The psychological values v_i^* of the consumption V_i of a subject i is defined by the relation

$$f_i\left(U_i + v_i^* V_i, 0, \ldots, W_i\right) = f\left(U_i, V_i, \ldots, W_i\right)$$

where $v_i^* V_i$ is the sum he would accept to receive to offset the drop in his consumption V_i to zero. The unit value v_i^* is generally much higher than the marginal value v_i corresponding to relations (17), (18) and (19).

In any event, a consumption is only advantageous when its psychological value is higher than its marginal value, because, if this were not so, it would be in the subject's interest to reduce his consumption V_i.

Conditions of Stable General Economic Equilibrium and Maximum Efficiency of the Economy

From condition (7) it follows that the necessary and sufficient condition for a situation $(\mathscr{E})$ to be of stable equilibrium and maximum efficiency is that the distributable surplus $\delta\sigma_u$ defined by (5) and (6) be negative or zero for every feasible modification $(\delta\mathscr{E})$, that is every modification that is compatible

with the constraint conditions, that is, the structural relations of the economy (2) and (4) above.

Condition (7) implies the two conditions (Allais 1943, p. 612)

$$\mathrm{d}\sigma_u = 0 \,(\text{first order condition}) \tag{21}$$

$$\mathrm{d}^2\sigma_u \leq 0 \,(\text{second order condition}) \tag{22}$$

for any realizable and reversible modification ($\delta\mathscr{E}$) in which the expressions of $\mathrm{d}\sigma_u$ and $\mathrm{d}^2\sigma_u$ represent the first and second differential of $\delta\sigma_u$.

Thus we have according to (21) and (22) using the above notations

$$\mathrm{d}\sigma_u = \sum_i \mathrm{d}\sigma_{ui} = \sum_i \mathrm{d}I_i / I'_{iu} = 0 \tag{23}$$

$$\mathrm{d}^2\sigma_u = \sum_i \frac{\overline{\mathrm{d}^2 f_{iu}}}{f'_{iu}} + \sum_j \frac{\overline{\mathrm{d}^2 f_{ju}}}{f'_{ju}} \leq 0 \quad \text{for} \quad \mathrm{d}\sigma_u = 0\,. \tag{24}$$

Actually, and according to relation (20), the first order condition (23) implies that when the quantities V_k are not nil, all the marginal values v_k are equal to a same value v and a same system of prices u, v,..., w *then exists for all the agents k concerned*, *such that.*

$$\frac{g'_{ku}}{u} = \frac{g'_{kv}}{v} = \cdots = \frac{g'_{kw}}{w}. \tag{25}$$

These equalities condense the general equimarginal principle into a single formulation. They express the fact that in a situation of equilibrium and maximum efficiency, the psychological (or objective) value v_k of the last dollar is the same, for any agent (consumption or production unit), whatever use it is put to.

For the quantities V_k which are nil (terminal equilibria), we necessarily have

$$v_k \leq v$$

since, if this were not true, the operator's interest would be to increase V_k from the value $V_k = 0$; he could indeed do this because of the existence of other operators who are in a situation of tangential equilibrium for good (V).

The second order condition (24) holds whether or not the $\mathrm{d}f_i$ are equal to zero. It is only subject to the constraint (21). If we consider only the modifications of the economy involving units k and l, condition (24) is written

$$\mathrm{d}^2\sigma_u = \frac{\overline{\mathrm{d}^2 f_{ku}}}{f'_{ku}} + \frac{\overline{\mathrm{d}^2 f_{lu}}}{f'_{lu}} \leq 0 \quad \text{for} \quad \mathrm{d}\sigma_{uk} + \mathrm{d}\sigma_{ul} = 0$$

shows that when in a situation of maximum efficiency consumption or production units consume (or produce) the same goods, one unit at most is in a situation of local concavity, that is, in a situation of marginal increasing returns (Allais 1968a, pp. 196–9; 1974a, n.125, p. 184; 1981, p. 65).

Consequently, when maximum efficiency obtains, most operators are in a situation of local convexity and marginal decreasing returns. However, this condition cannot be interpreted as meaning that all fields of choice and production are convex everywhere, this hypothesis being totally contradicted by observed data.

When local convexity obtains for a consumption unit, its index of preference is effectively at a maximum, subject to the budgetary constraint, equilibrium prices being taken as given. Similarly, if local convexity obtains for a production unit, the unit's income is effectively at a maximum, equilibrium prices again being taken as given. However, these two principles, which in any case could be valid only for a situation of maximum efficiency, cannot be considered as corresponding in all cases to optimum behaviour, and they cannot be taken to be of general value. As a matter of fact and for instance, if, in a situation of maximum efficiency, a production unit is in a situation of local concavity, its income is minimum, the equilibrium prices being considered as given.

Conditions (25) and (24) show the total symmetry of the implications of the psychological and technical structures of the economy.

Approximate Value of the Economic Loss Corresponding to the Non-equality of Marginal Values in the Neighbourhood of a Situation of Maximum Efficiency

The integration of Eq. (20) along a path leading to a state of maximum efficiency leads to the following approximate estimate to within third order accuracy of the global loss involved in the initial situation (relation 8)

$$\sigma_u^* \sim \frac{1}{2u} \sum_{v} \sum_{\substack{k,l \\ k<l}} (v_k - v_l)\delta V_{kl}^*. \qquad (26)$$

In this relation, the quantities $v_k - v_l$ represent the differences of marginal values in the initial state considered, and the δV_{kl}^* are the quantities of the good (V) received by operator k from operator l in the transition from the initial to the final state. Relation (26) is of the broadest generality, and holds whatever the initial state (Allais 1952a, pp. 31–2, n. 8; 1968a, vol. 2, p. 207; 1981, p. 110).

Its simplicity is really extraordinary in view of the complexity of the concept it represents, namely the maximum of the distributable surplus for all the modifications which the economy can undergo while leaving the preference indexes unchanged.

In the neighbourhood of a situation of maximum efficiency, the $(v_k - v_l)$ and δV_{kl}^* are of the first order quantities, where as the loss σ_u^* is only of the second order. However, since the δV_{kl}^* are of the first order, the variations δI_i of the preference indexes are also of the first order. As a result, and for instance, in the neighbourhood of a situation of maximum efficiency, taxes have major first order effects on the distribution of income but only second order effects on the efficiency of the economy.

On the theoretical foundations of the equimarginal principle, see Allais (1943, pp. 604–56, 1945, 1952a, pp. 28–32; 1967, 1968a, vol. 2, 1971, 1973a, 1973b, 1974a, 1974b, 1981 and 1986). Illustrative models: Allais (1943, Annexe I, pp. 4–24; 1945, pp. 57–69). On its extension see: cases of perfect and imperfect foresight: Allais (1943, pp. 343–84; 1947, pp. 23–228; 1964, 1967, 1968a, vol. 2). Illustrative models: (1947, pp. 631–771). Capitalistic optimum theory: Allais (1947, pp. 179–228; 1962, 1963). Demographic optimum theory: Allais (1943, pp. 749–85). Case of risk: Allais (1952b). Application of marginal analysis to transport: Allais (1964 and 1987). For a general overview on the meaning, limits, generalizations, and history of the equimarginal analysis see Allais (1987).

General Overview

Theory of Surpluses and Marginal Analysis

As a matter of fact a single relation, the relation (20) (or the equivalent relation (16)) condenses the whole marginal approach as it has developed for over a century. Subject only to the hypotheses of continuity and derivability implied by any marginal theory, it applies in all cases, and its simplicity is really extraordinary.

It also shows that equilibrium and maximum efficiency can obtain only when all marginal values are equal, which is the equimarginal principle.

The equimarginal principle was discovered first by Gossen (1854), and rediscovered, broadened and introduced independently into economics by Jevons (1871), Menger (1871) and Walras (1874–7). In the following years numerous new developments of the principle have been presented by their immediate successors, especially by Edgeworth (1881), Irving Fisher (1892) and Vilfredo Pareto (1896–1911). Particularly striking illustrations of the role of differences in marginal equivalences are Ricardo's theory of comparative costs (1817) and Dupuit's theory of economic losses (1844–53).

This principle corresponds to the outcome of the dynamic process of the economy induced by differences in marginal equivalences. According to Irving Fisher (1892), with whose judgement I agree fully, 'No idea has been more fruitful in the history of economic science.' Its applications

and generalizations dominate all economic analysis in real terms.

From the foregoing a double conclusion emerges: the classical theory of marginal equivalences is irreplaceable to make understandable the underlying nature of all economic phenomena; the general theory of surplus, of which classical marginal theory is only a special case, allows one to extend the propositions of marginal analysis to the most general case of discrete variations and indivisibilities.

As important as the analysis of the conditions of general equilibrium and maximum efficiency may be, the analysis of the dynamic processes which enable surpluses to be generated from a given situation is much more important. From this point of view the analyses by Dupuit, Jevons, Edgeworth, Pareto, and the marginal school and its predecessors in general, appear much more realistic than the contributions which rest only upon the consideration of Walras's general model of equilibrium.

In fact, what is really important is not so much the knowledge of the properties of a state of maximum efficiency as the rules of the game which have been applied to the economy effectively to move nearer to a state of maximum efficiency.

The decentralized search for surpluses is truly the dynamic principle from which a thorough and yet very simple conception of the operation of the whole economy can be derived. Whereas in the market economy model the search for efficiency is essentially focused on the determination of a certain set of prices, the analysis of the model of the economy of markets is based on the search for potential surpluses and their realization. Not only is the economy of markets model much more realistic than the market economy model while lending itself to much simpler proofs, but also these proofs are not subordinated to any restrictive assumptions relating to continuity, differentiability of functions, or convexity. All of economic dynamics is reduced to a single principle: the search for and realization of potential surpluses, which leads to the minimization of loss for the economy as a whole.

On all these points see especially: Allais (1971 and 1974a).

The Tendencies of the Contemporary Literature

From Walras on, the literature became progressively – and unduly – concentrated on equilibrium analysis which, however interesting it could be, is less so than the analysis of the processes by which the economy tends at any time towards situations of equilibrium which in fact are never reached.

Today there is a tendency to neglect the dynamic marginal approach based on the consideration of differences in marginal equivalences; and in the name of a so-called rigour it has been replaced by new theories. A fortiori, the general theory of surpluses which generalizes marginal analysis is simply ignored. This development, which in reality, and despite the too-widely held belief to the contrary, represents an immense step backward, basically stems from the unquestioning acceptance of 'established truths' taught by the dominant 'establishments', whose only real basis is their incessant repetition.

As a matter of fact the guiding principles of the contemporary theories descending from Walras: the adoption of the market economy model; the hypothesis that a common price system applicable to all operators prevails at each instant; the assumption of general convexity; and the exaltation of mathematical formalism of the theory of sets to the detriment of conformity with actual facts, constitute an impediment to any genuine progress in analysis of the economy in real terms.

The essential difference between the market economy model and the model of the economy of markets is that, in the latter, the exchanges leading to equilibrium take place successively at different prices, and that, at any given moment, the price sets used by different operators are not necessarily the same. Whereas in the first model the final situation is determined totally by the initial situation, which correspondingly plays a privileged role without any real justification, in the second the final situation depends both on the

initial situation and the path taken from it to the final situation (Fig. 1).

Whereas the market economy model postulates perfect competition and a large number, if not an infinity, of operators, the model of the economy of markets applies just as well to the cases of monopoly as to the cases of competition.

Not only is the market economy model unrealistic, but it also gives rise to considerable mathematical difficulties when an attempt is made to demonstrate the above three fundamental theorems. Whether differential calculus or set theory is used, the theorems can only be demonstrated under extremely restrictive conditions, and the difficulties they imply are, from an economic standpoint, completely artificial, for they arise solely from the unrealistic nature of the model used. Paradoxically, whereas these restrictive assumptions are totally unrealistic, most of the theoretical difficulties encountered disappear, as shown above, once they are discarded.

The market economy approach leads to imposing on any economic model, for it to be considered satisfactory, conditions which actually apply to a particular model, which are generally not fulfilled in reality, and for which, at all events, no rigorous justification can be found.

By departing from the great tradition of marginal theory and by adopting an unrealistic model and unrealistic assumptions, the contemporary theories, purely mathematical, have doomed themselves to sterility as regards the understanding of reality.

On the contemporary theories see especially: Samuelson (1947), Arrow (1968), Debreu (1959 and 1985), Blaug (1979 and 1985), Arrow and Hahn (1971), Hutchison (1977, pp. 62–97 and 161–70), Woo (1985), and Allais (1952b, 1968b, 1968e, 1971, 1974a, and 1981).

See Also

- ▶ Allais, Maurice (Born 1911)
- ▶ Efficient Allocation
- ▶ General Equilibrium
- ▶ Optimality and Efficiency
- ▶ Surplus

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique. Première partie. I'économie pure*, 2 vols. Paris: Ateliers Industria. 2nd ed., *Traité d'économie pure*, 5 vols, Paris: Imprimerie Nationale, Paris, 1952. (The second edition is identical to the first, except for a new Introduction.)

Allais, M. 1945. *Economie pure et rendement social*. Paris: Sirey.

Allais, M. 1947. *Economie et intérêt*, 2 vols. Paris: Imprimerie Nationale and Librairie des Publications Officielles.

Allais, M. 1952a. *Introduction to the 2nd edn of Allais (1943)*. Paris: Imprimerie Nationale.

Allais, M. 1952b. L'éxtension des théories de l'équilibre économique général et du rendement social au cas du risque. *Colloques Internationaux du Centre National de la Recherche Scientifique* 40, *Econométrie*, 81–120. A summarized version was published under the same title in *Econometrica*, (1953), 269–290.

Allais, M. 1962. The influence of the capital output ratio on real national income. *Econometrica* 30: 700–728. Republished in American Economic Association, *Readings in Welfare Economics*, vol. 12, with an additional Note, 1969.

Allais, M. 1963. The role of capital in economic development. In *Study work on the econometric approach to development planning*, Pontificiae Academiae Scientiarum Scripta Varia 28, Pontifica Academia Scientiarum, Amsterdam: North-Holland, 1963 and Chicago: Rand McNally, 1965.

Allais, M. 1964. La theorie economique et la tarification optimum de l'usage des infrastructures de transport. *La Jaune et la Rouge* (publication of the Société Amicale des Anciens Eléves de l'Ecole Polytechnique), special issue *Les Transports*, Paris, 1964.

Allais, M. 1967. Les conditions de l'éfficacite dans l'économie. Fourth International Seminar, Centro Studi e Ricerche su Problemi Economico-Sociali, Milan. Italian translation: 'Le condizioni dell' efficienza nell' economia' in *Programmazione E Progresso Economico*. Milan: Franco Angeli, 1969. Original French text in M. Allais, *Les Fondements du Calcul Economique*, vol. 1. Paris: Ecole Nationale Supérieure des Mines de Paris, 1967.

Allais, M. 1968a. *Les fondements du calcul économique*, 3 vols. Paris: Ecole Nationale Supérieure des Mines, Paris, vol. 1, 1967, and vols 2 and 3, 1968.

Allais, M. 1968b. The conditions of efficiency in the economy. *Economia Internazionale* 21 (3): 399–420.

E

Allais, M. 1968c. Pareto, Vilfredo: Contributions to economics. In *International encyclopedia of the social sciences*, vol. 2. New York: Macmillan and Free Press.

Allais, M. 1968d. Fisher, Irving. In *International encyclopedia of the social sciences*, vol. 5. New York: Macmillan and Free Press.

Allais, M. 1968e. L'économique en tant que science. *Revue d'Economie Politique*, January–February, 5–30. Trans. as 'Economics as a science', *Cahiers Vilfredo Pareto*, (1968), 5–24.

Allais, M. 1971. Les théories de l'équilibre economique général et de l'éfficacité maximale – impasse récentes et nouvelles perspectives. Congrés des Economistes de Langue Francçaise, 2–6 June. *Revue d'Economie Politique* 3: 331–409. Spanish translation: 'Las theorias del equilibrio economico general y de la eficacia maxima – recientes callejones sin salida y nuevas perspectivas'. *El Trimestre Economico* (Mexico), 39(1972), 557–633; English translation: see Allais (1974a).

Allais, M. 1973a. La théorie gélnérale des surplus et l'apport fondamental de Vilfredo Pareto. *Revue d'Economie Politique* 6: 1044–1097.

Allais, M. 1973b. The general theory of surplus and Pareto's fundamental contribution. *Convegno Internazionale Vilfredo Pareto*. Roma, 25–27 October, Rome: Accademia Nazionale dei Lincei, 1975 (English trans. of Allais, 1973a.)

Allais, M. 1974a. Theories of general economic equilibrium and maximum efficiency. Vienna Institute for Advanced Studies. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwödiauer. Dordrecht: Reidel, 1977. (English version of Allais, 1971, with some additions.)

Allais, M. 1974b. Les implications de rendements croissants et décroissants sur les conditions de I'équilibre economique général et d'uned éfficacité maximale. In *Hommage à François Perroux*. Grenoble: Presses Universitaires de Grenoble, 1978.

Allais, M. 1981. *La théorie générale des surplus. Economies et Sociétés*, 2 vols. Montrouge institut de sciences mathématique et économies appliquées.

Allais, M. 1985. The concepts of surplus and loss and the reformulation of the theories of stable general economic equilibrium and maximum efficiency. In *Foundations and dynamics of economic knowledge*, ed. M. Baranzini and R. Scazzieri. Oxford: Basil Blackwell.

Allais, M. 1987. The equimarginal principle, meaning, limits, and generalisations. *Centre d'Analyse Economique. Revista internazionale di scienze economiste e commerciale.*

Arrow, K.J. 1968. Economic equilibrium. In *International encyclopedia of the social sciences*, vol. 4. New York: Macmillan/Free Press.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco/Edinburgh: Holden-Day/ Oliver.

Blaug, M. 1979. *Economic theory in retrospect*. 4th ed. London: Heinemann Educational Books, 1985.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1985. *Theoretic models: Mathematical form and economic content*. Frisch Memorial Lecture, Fifth World Congress of the Econometric Society, MIT, 17–24 August.

Dupuit, J. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées*, 2nd series, Mémoires et Documents No. 116, vol. 8.

Dupuit, J. 1849. De l'influence des péages sur l'utilité des voies de communication. *Annales des Ponts et Chaussées*, 2nd series.

Dupuit, J. 1853. De l'utilite et de sa mesure. *Journal des Economistes* 36 (147): 1–28.

Edgeworth, F.Y. 1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. London: Kegan Paul. Reprinted, New York: Kelley, 1953.

Fisher, I. 1892. *Mathematical investigations in the theory of value and prices*. New Haven: Yale University Press, 1925.

Gossen, H.H. von. 1854. *Entwickelung der Gesetze des menschlichen Verkehrs und der daraus fliessender Regeln für menschliches Handeln.* 3rd ed., introduction by Friedrich Hayek, Berlin: Präger, 1927.

Hutchison, T.W. 1977. *Knowledge and ignorance in economics*. Oxford: Blackwell.

Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan. 5th ed., trans. as *La théorie de économie politique*. Paris: Giard, 1909.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Vienna: Braumneller.

Pareto, V. 1896–7. *Cours d'économie politique*, 2 vols. Lausanne: Rougé. Reprinted, Geneva: Droz, 1964.

Pareto, V. 1901. Anwendungen der Mathematik auf Nationalökonomie. *Encyklopädie der Mathematichen Wissenschaften*, vol. 1. Leipzig.

Pareto, V. 1906. *Manuale d'economia politica*. Milan. Trans. as *Manuel d'économie politique*. Paris: Giard et Brière, 1909, and Geneva: Droz, 1966, and as *Manual of political economy*. Reprinted, New York: Kelley, 1971.

Pareto, V. 1911. Economie mathématique. *Encyclopedie des Sciences Mathématiques*. Paris: Gauthier–Villars, 1911, also in *Statistique et économie mathématique*. Geneva: Droz, 1966, published in English as 'Mathematical economics', *International economic papers no*. 5. New York, 1955.

Ricardo, D. 1817. On the principles of political economy and taxation. Vol. I of *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951–1955.

Samuelson, P.A. 1947. *Foundations of economic analysis*. 2nd ed. Cambridge, MA: Harvard University Press, 1948.

Samuelson, P.A. 1950. Evaluation of real national income. *Oxford Economic Papers NS* 2: 1–29.

Walras, L. 1874–7. *Eléments d'économie politique pure – théorie de la richesse sociale*. 6th ed. Paris: Guillaumin; reprinted, Paris: Pichon et Durand-Auzias,

1952. English trans. of the 6th ed., as *Elements of pure economics*, ed. W. Jaffé, London: Allen & Unwin, 1954.
Woo, H.K.H. 1985. *What's wrong with formalization in economics? – An epistemological critique*. Hong Kong: Hong Kong Institute of Economic Science.

Economic Theory and the Hypothesis of Rationality

Kenneth J. Arrow

In this paper, I want to disentangle some of the senses in which the hypothesis of rationality is used in economic theory. In particular, I want to stress that rationality is not a property of the individual alone, although it is usually presented that way. Rather, it gathers not only its force but also its very meaning from the social context in which it is embedded. It is most plausible under very ideal conditions. When these conditions cease to hold, the rationality assumptions become strained and possibly even self-contradictory. They certainly imply an ability at information processing and calculation that is far beyond the feasible and that cannot well be justified as the result of learning and adaptation.

Let me dismiss a point of view that is perhaps not always articulated but seems implicit in many writings. It seems to be asserted that a theory of the economy must be based on rationality, as a matter of principle. Otherwise, there can be no theory. This position has even been maintained by some who accept that economic behaviour is not completely rational. John Stuart Mill (1848, bk. 2, ch. 4) argued that custom, not competition, governs much of the economic world. But he adds that the only possible theory is that based on competition (which, in his theories, includes certain elements of rationality, particularly shifting capital and labour to activities that yield higher returns); 'Only through the principle of competition has political economy any pretension to the character of science', ([1848] 1909, p. 242).

Certainly, there is no general principle that prevents the creation of an economic theory based on hypotheses other than that of rationality. There are indeed some conditions that must be laid down for an acceptable theoretical analysis of the economy. Most centrally, it must include a theory of market interactions, corresponding to market clearing in the neoclassical general equilibrium theory. But as far as individual behaviour is concerned, any coherent theory of reactions to the stimuli appropriate in an economic context (prices in the simplest case) could in principle lead to a theory of the economy. In the case of consumer demand, the budget constraint must be satisfied but many theories can easily be devised that are quite different from utility maximization. For example, habit formation can be made into a theory; for a given price-income change, choose the bundle that satisfies the budget constraint and that requires the least change (in some suitably defined sense) from the previous consumption bundle. Though there is an optimization in this theory, it is different from utility maximization; for example, if prices and income return to their initial levels after several alterations, the final bundle purchased will not be the same as the initial. This theory would strike many lay observers as plausible, yet it is not rational as economists have used that term. Without belabouring the point, I simply observe that this theory is not only a logically complete explanation of behaviour but one that is more powerful than standard theory and at least as capable of being tested.

Not only is it possible to devise complete models of the economy on hypotheses other than rationality, but in fact virtually every practical theory of macroeconomics is partly so based. The price- and wage-rigidity elements of Keynesian theory are hard to fit into a rational framework, though some valiant efforts have been made. In the original form, the multiplier was derived from a consumption function depending only on current income. Theories more nearly based on rationality make consumption depend on lifetime or 'permanent' income and reduce the magnitude of the multiplier and, with it, the explanatory power of the Keynesian model. But if the Keynesian model

is a natural target of criticism by the upholders of universal rationality, it must be added that monetarism is no better. I know of no serious derivation of the demand for money from a rational optimization. The loose arguments that substitute for a true derivation, Friedman's economizing on shoe leather or Tobin's transaction demand based on costs of buying and selling bonds, introduce assumptions incompatible with the costless markets otherwise assumed. The use of rationality in these arguments is ritualistic, not essential. Further, the arguments used would not suggest a very stable relation but rather one that would change quickly with any of the considerable changes in the structure and technology of finance. Yet the stability of the demand function for money must be essential to any form of monetarism, not excluding those rational expectations models in which the quantity theory plays a major role.

I believe that similar observations can be made about a great many other areas of applied economics. Rationality hypotheses are partial and frequently, if not always, supplemented by assumptions of a different character.

So far, I have argued simply that rationality is not in principle essential to a theory of the economy, and, in fact, theories with direct application usually use assumptions of a different nature. This was simply to clear the ground so that we can discuss the role of rationality in economic theory. As remarked earlier, rationality in application is not merely a property of the individual. Its useful and powerful implications derive from the conjunction of individual rationality and the other basic concepts of neoclassical theory – equilibrium, competition, and completeness of markets. The importance of all these assumptions was first made explicit by Frank Knight (1921, pp. 76–79). In the terms of Knight's one-time student, Edward Chamberlin (1950, pp. 6–7), we need not merely pure but perfect competition before the rationality hypotheses have their full power.

It is largely this theme on which I will expand. When these assumptions fail, the very concept of rationality becomes threatened, because perceptions of others and, in particular, of their rationality become part of one's own rationality. Even if there is a consistent meaning, it will involve computational and informational demands totally at variance with the traditional economic theorist's view of the decentralized economy.

Let me add one parenthetic remark to this section. Even if we make all the structural assumptions needed for perfect competition (whatever is needed by way of knowledge, concavity in production, absence of sufficient size to create market power, etc.), a question remains. How can equilibrium be established? The attainment of equilibrium requires a disequilibrium process. What does rational behaviour mean in the presence of disequilibrium? Do individuals speculate on the equilibrating process? If they do, can the disequilibrium be regarded as, in some sense, a higher-order equilibrium process? Since no one has market power, no one sets prices; yet they are set and changed. There are no good answers to these questions, and I do not pursue them. But they do illustrate the conceptual difficulties of rationality in a multiperson world.

Rationality as Maximization in the History of Economic Thought

Economic theory, since it has been systematic, has been based on some notion of rationality. Among the classical economists, such as Smith and Ricardo, rationality had the limited meaning of preferring more to less; capitalists choose to invest in the industry yielding the highest rate of return, landlords rent their property to the highest bidder, while no one pays for land more than it is worth in product. Scattered remarks about technological substitution, particularly in Ricardo, can be interpreted as taking for granted that, in a competitive environment, firms choose factor proportions, when they are variable, so as to minimize unit costs. To be generous about it, their rationality hypothesis was the maximization of profits by the firm, although this formulation was not explicitly achieved in full generality until the 1880s.

There is no hypothesis of rationality on the side of consumers among the classicists. Not until John Stuart Mill did any of the English classical economists even recognize the idea that demand

might depend on price. Cournot had the concept a bit earlier, but neither Mill nor Cournot noticed – although it is obvious from the budget constraint alone – that the demand for any commodity must depend on the price of all commodities. That insight remained for the great pioneers of the marginalist revolution, Jevons, Walras, and Menger (anticipated, to be sure, by the Gregor Mendel of economics, H.H. Gossen, whose major work, completely unnoticed at the time of publication [1854], has now been translated into English [1983]). Their rationality hypothesis for the consumer was the maximization of the utility under a budget constraint. With this formulation, the definition of demand as a function of all prices was an immediate implication, and it became possible to formulate the general equilibrium of the economy.

The main points in the further development of the utility theory of the consumer are well known. (1) Rational behaviour is an ordinal property. (2) The assumption that an individual is behaving rationally has indeed some observable implications, the Slutsky relations, but without further assumptions, they are not very strong. (3) In the aggregate, the hypothesis of rational behaviour has in general noimplications; that is, for any set of aggregate excess demand functions, there is a choice of preference maps and of initial endowments, one for each individual in the economy, whose maximization implies the given aggregate excess demand functions (Sonnenschein 1973; Mantel 1974; Debreu 1974; for a survey, see Shafer and Sonnenschein 1982, sec. 4).

The implications of the last two remarks are in contradiction to the very large bodies of empirical and theoretical research, which draw powerful implications from utility maximization for, respectively, the behaviour of individuals, most especially in the field of labour supply, and the performance of the macroeconomy based on 'new classical' or 'rational expectations' models. In both domains, this power is obtained by adding strong supplementary assumptions to the general model of rationality. Most prevalent of all is the assumption that all individuals have the same utility function (or at least that they differ only in broad categories based on observable magnitudes, such as family size). But this postulate leads to curious and, to my mind, serious difficulties in the interpretation of evidence. Consider the simplest models of human capital formation. Cross-sectional evidence shows an increase of wages with education or experience, and this is interpreted as a return on investment in the form of foregone income and other costs. But if all individuals are alike, why do they not make the same choice? Why do we observe a dispersion? In the human capital model (a particular application of the rationality hypothesis), the only explanation must be that individuals are not alike, either in ability or in tastes. But in that case the cross-sectional evidence is telling us about an inextricable mixture of individual differences and productivity effects. Analogously, in macroeconomic models involving durable assets, especially securities, the assumption of homogeneous agents implies that there will never be any trading, though there will be changes in prices.

This dilemma is intrinsic. If agents are all alike, there is really no room for trade. The very basis of economic analysis, from Smith on, is the existence of differences in agents. But if agents are different in unspecifiable ways, then remark (3) above shows that very few, if any, inferences can be made. This problem, incidentally, already exists in Smith's discussion of wage differences. Smith did not believe in intrinsic differences in ability; a porter resembled a philosopher more than a greyhound did a mastiff. Wage differences then depended on the disutilities of different kinds of labour, including the differential riskiness of income. This is fair enough and insightful. But, if taken seriously, it implies that individuals are indifferent among occupations, with wages compensating for other differences. While there is no logical problem, the contradiction to the most obvious evidence is too blatant even for a rough approximation.

I have not carried out a scientific survey of the uses of the rationality hypothesis in particular applications. But I have read enough to be convinced that its apparent force comes only from the addition of supplementary hypotheses. Homogeneity across individual agents is not the only auxiliary assumption, though it is the deepest. Many

assumptions of separability are frequently added. Indeed, it has become a working methodology to start with very strong assumptions of additivity and separability, together with a very short list of relevant variables, to add others only as the original hypotheses are shown to be inadequate, and to stop when some kind of satisfactory fit is obtained. A failure of the model is attributed to a hither to overlooked benefit or cost. From a statistical viewpoint, this stopping rule has obvious biases. I was taught as a graduate student that data mining was a major crime; morality has changed here as elsewhere in society, but I am not persuaded that all these changes are for the better.

The lesson is that the rationality hypothesis is by itself weak. To make it useful, the researcher is tempted into some strong assumptions. In particular, the homogeneity assumption seems to me to be especially dangerous. It denies the fundamental assumption of the economy, that it is built on gains from trading arising from individual differences. Further, it takes attention away from a very important aspect of the economy, namely, the effects of the distribution of income and of other individual characteristics on the workings of the economy. To take a major example, virtually all of the literature on savings behaviour based on aggregate data assumes homogeneity. Yet there have been repeated studies that suggest that saving is not proportional to income, from which it would follow that distributional considerations matter. (In general, as data have improved, it has become increasingly difficult to find any simple rationally based model that will explain savings, wealth, and bequest data.)

The history of economic thought shows some other examples and difficulties with the application of the rationality hypothesis. Smith and the later classicists make repeated but unelaborated references to risk as a component in wage differences and in the rate of return on capital (e.g., Mill [1848] 1909, pp. 385, 406, 407, 409). The English marginalists were aware of Bernoulli's expected-utility theory of behaviour under uncertainty (probably from Todhunter's *History of the Theory of Probability*) but used it only in a qualitative and gingerly way (Jevons [1871] 1965, pp. 159–60; Marshall 1920, pp. 842–3). It was really not until the last 30 years that it has been used systematically as an economic explanation, and indeed its use coincided with the first experimental evidence against it (see Allais 1979). The expected-utility hypothesis is an interesting transition to the theme of the next section. It is in fact a stronger hypothesis than mere maximization. As such it is more easily tested, and it leads to stronger and more interesting conclusions. So much, however, has already been written about this area that I will not pursue it further here.

Rationality, Knowledge, and Market Power

It is noteworthy that the everyday usage of the term 'rationality' does not correspond to the economist's definition as transitivity and completeness, that is, maximization of something. The common understanding is instead the complete exploitation of information, sound reasoning, and so forth. This theme has been systematically explored in economic analysis, theoretical and empirical, only in the last 35 years or so. An important but neglected predecessor was Holbrook Working's random-walk theory of fluctuations in commodity futures and securities prices (1953). It was based on the hypothesis that individuals would make rational inferences from data and act on them; specifically, predictability of future asset prices would be uncovered and used as a basis for current demands, which would alter current prices until the opportunity for gain was wiped out.

Actually, the classical view had much to say about the role of knowledge, but in a very specific way. It emphasized how a complete price system would require individuals to know very little about the economy other than their own private domain of production and consumption. The profoundest observation of Smith was that the system works behind the backs of the participants; the directing 'hand' is 'invisible'. Implicitly, the acquisition of knowledge was taken to be costly.

Even in a competitive world, the individual agent has to know all (or at least a great many)

prices and then perform an optimization based on that knowledge. All knowledge is costly, even the knowledge of prices. Search theory, following Stigler (1961), recognized this problem. But search theory cannot easily be reconciled with equilibrium or even with individual rationality by price setters, for identically situated sellers should set identical prices, in which case there is nothing to search for.

The knowledge requirements of the decision may change radically under monopoly or other forms of imperfect competition. Consider the simplest case, pure monopoly in a one-commodity partial equilibrium model, as originally studied by Cournot in Cournot 1838. The firm has to know not only prices but a demand curve. Whatever definition is given to complexity of knowledge, a demand curve is more complex than a price. It involves knowing about the behaviour of others. Measuring a demand curve is usually thought of as a job for an econometrician. We have the curious situation that scientific analysis imputes scientific behaviour to its subjects. This need not be a contradiction, but it does seem to lead to an infinite regress.

From a general equilibrium point of view, the difficulties are compounded. The demand curve relevant to the monopolist must be understood *mutatis mutandis*, not *ceteris paribus*. A change in the monopolist's price will in general cause a shift in the purchaser's demands for other goods and therefore in the prices of those commodities. These price changes will in turn affect by more than one channel the demand for the monopolist's produce and possibly also the factor prices that the monopolist pays. The monopolist, even in the simple case where there is just one in the entire economy, has to understand all these repercussions. In short, the monopolist has to have a full general equilibrium model of the economy.

The informational and computational demands become much stronger in the case of oligopoly or any other system of economic relations where at least some agents have power against each other. There is a qualitatively new aspect to the nature of knowledge, since each agent is assuming the *rationality* of other agents. Indeed, to construct a rationality-based theory of economic behaviour, even more must be assumed, namely, that the rationality of all agents must be *common knowledge*, to use the term introduced by the philosopher David Lewis (1969). Each agent must not only know that the other agents (at least those with significant power) are rational but know that each other agent knows every other agent is rational, know that every other agent knows that every other agent is rational, and so forth (see also Aumann 1976). It is in this sense that rationality and the knowledge of rationality is a social and not only an individual phenomenon.

Oligopoly is merely the most conspicuous example. Logically, the same problem arises if there are two monopolies in different markets. From a practical viewpoint, the second case might not offer such difficulties if the links between the markets were sufficiently loose and the monopolies sufficiently small on the scale of the economy that interaction was negligible; but the interaction can never be zero and may be important. As usually presented, bargaining to reach the contract curve would, in the simplest case, require common knowledge of the bargainer's preferences and production functions. It should be obvious how vastly these knowledge requirements exceed those required for the price system. The classic economists were quite right in emphasizing the importance of limited knowledge. If every agent has a complete model of the economy, the hand running the economy is very visible indeed.

Indeed, under these knowledge conditions, the superiority of the market over centralized planning disappears. Each individual agent is in effect using as much information as would be required for a central planner. This argument shows the severe limitations in the argument that property rights suffice for social rationality even in the absence of a competitive system (Coase 1960).

One can, as many writers have, discuss bargaining when individuals have limited knowledge of each other's utilities (similarly, we can have oligopoly theory with limited knowledge of the cost functions of others: see, e.g., Arrow 1979). Oddly enough, it is not clear that limited knowledge means a smaller quantity of information than complete knowledge, and optimization

under limited knowledge is certainly computationally more difficult. If individuals have private information, the others form some kind of conjecture about it. These conjectures must be common knowledge for there to be a rationality-based hypothesis. This seems to have as much informational content and to be as unlikely as knowing the private information. Further, the optimization problem for each individual based on conjectures (in a rational world, these are probability distributions) on the private information of others is clearly a more difficult and therefore computationally more demanding problem than optimization when there is no private information.

Rational Knowledge and Incomplete Markets

It may be supposed from the foregoing that informational demands are much less in a competitive world. But now I want to exemplify the theme that perfect, not merely pure, competition is needed for that conclusion and that perfect competition is a stronger criterion than Chamberlin perhaps intended. A complete general equilibrium system, as in Debreu (1959), requires markets for all contingencies in all future periods. Such a system could not exist. First, the number of prices would be so great that search would become an insuperable obstacle; that is, the value of knowing prices of less consequence, those of events remote in time or of low probability, would be less than the cost so that these markets could not come into being. Second, markets conditional on privately observed events cannot exist by definition.

In any case, we certainly know that many – in fact, most – markets do not exist. When a market does not exist, there is a gap in the information relevant to an individual's decision, and it must be filled by some kind of conjecture, just as in the case of market power. Indeed, there turn out to be strong analogies between market power and incomplete markets, though they seem to be very different phenomena.

Let me illustrate with the rational expectations equilibrium. Because of intertemporal relations in consumption and production, decisions made today have consequences that are anticipated. Marshall (1920, bk 5, chs 3–5) was perhaps the first economist to take this issue seriously. He introduced for this purpose the vague and muddled concepts of the short and long runs, but at least he recognized the difficulties involved, namely, that some of the relevant terms of trade are not observable on the market. (Almost all other accounts implicitly or explicitly assumed a stationary state, in which case the relative prices in the future and between present and future are in effect current information. Walras (1874, lessons 23–25) claimed to treat a progressive state with net capital accumulation, but he wound up unwittingly in a contradiction, as John Eatwell has observed in an unpublished dissertation. Walras's arguments can only be rescued by assuming a stationary state.) Marshall in effect made current decisions, including investment and savings, depend on expectations of the future. But the expectations were not completely arbitrary; in the absence of disturbances, they would converge to correct values. Hicks (1946, chs 9–10) made the dependence of current decisions on expectations more explicit, but he had less to say about their ultimate agreement with reality.

As has already been remarked, the full competitive model of general equilibrium includes markets for all future goods and, to take care of uncertainty, for all future contingencies. Not all of these markets exist. The new theoretical paradigm of rational expectations holds that each individual forms expectations of the future on the basis of a correct model of the economy, in fact, the same model that the econometrician is using. In a competitive market-clearing world, the individual agent needs expectations of prices only, not of quantities. For a convenient compendium of the basic literature on rational expectations, see Lucas and Sargent (1981). Since the world is uncertain, the expectations take the form of probability distributions, and each agent's expectations are conditional on the information available to him or her.

As can be seen, the knowledge situation is much the same as with market power. Each agent has to have a model of the entire economy to preserve rationality. The cost of knowledge, so emphasized by the defenders of the price system

as against centralized planning, has disappeared; each agent is engaged in very extensive information gathering and data processing.

Rational expectations theory is a stochastic form of perfect foresight. Not only the feasibility but even the logical consistency of this hypothesis was attacked long ago by Morgenstern (1935). Similarly, the sociologist Robert K. Merton (1957) argued that forecasts could be self-denying or self-fulfilling; that is, the existence of the forecast would alter behaviour so as to cause the forecast to be false (or possibly to make an otherwise false forecast true). The logical problems were addressed by Grunberg and Modigliani (1954) and by Simon (1957, ch. 5). They argued that, in Merton's terms, there always existed a self-fulfilling prophecy. If behaviour varied continuously with forecasts and the future realization were a continuous function of behaviour, there would exist a forecast that would cause itself to become true. From this argument, it would appear that the possibility of rational expectations cannot be denied. But they require not only extensive first-order knowledge but also common knowledge, since predictions of the future depend on other individuals' predictions of the future. In addition to the information requirements, it must be observed that the computation of fixed points is intrinsically more complex than optimizing.

Consider now the signalling equilibrium originally studied by Spence (1974). We have large numbers of employers and workers with free entry. There is no market power as usually understood. The ability of each worker is private information, known to the worker but not to the employer. Each worker can acquire education, which is publicly observable. However, the cost of acquiring the education is an increasing function of ability. It appears natural to study a competitive equilibrium. This takes the form of a wage for each educational level, taken as given by both employers and workers. The worker, seeing how wages vary with education, chooses the optimal level of education. The employer's optimization leads to an 'informational equilibrium' condition, namely, that employers learn the average productivity of workers with a given educational level. What dynamic process would lead the market to learn these productivities is not clear, when employers are assumed unable to observe the productivity of individual workers. There is more than one qualitative possibility for the nature of the equilibrium. One possibility, indeed, is that there is no education, and each worker receives the average productivity of all workers (I am assuming for simplicity that competition among employers produces a zero-profit equilibrium). Another possibility, however, is a dispersion of workers across educational levels; it will be seen that in fact workers of a given ability all choose the same educational level, so the ability of the workers could be deduced from the educational level ex post.

Attractive as this model is for certain circumstances, there are difficulties with its implementation, and at several different levels. (1) It has already been noted that the condition that, for each educational level, wages equal average productivity of workers is informationally severe. (2) Not only is the equilibrium not unique, but there is a continuum of possible equilibria. Roughly speaking, all that matters for the motivation of workers to buy education are the relative wages at different educational levels; hence, different relations between wages and education are equally self-fulfilling. As will be seen below, this phenomenon is not peculiar to this model. On the contrary, the existence of a continuum of equilibria seems to be characteristic of many models with incomplete markets. Extensive non-uniqueness in this sense means that the theory has relatively little power. (3) The competitive equilibrium is fragile with respect to individual actions. That is, even though the data of the problem do not indicate any market power, at equilibrium it will frequently be possible for any firm to profit by departing from the equilibrium.

Specifically, given an equilibrium relation between wages and education, it can pay a firm to offer a different schedule and thereby make a positive profit (Riley 1979). This is not true in a competitive equilibrium with complete markets, where it would never pay a firm to offer any price or system of prices other than the market's. So far, this instability of competitive equilibrium is a

property peculiar to signalling models, but it may be more general.

As remarked above, the existence of a continuum of equilibria is now understood to be a fairly common property of models of rational market behaviour with incomplete information. Thus, if there were only two commodities involved and therefore only one price ratio, a continuum of equilibria would take the form of a whole interval of price ratios. This multiplicity would be non-trivial, in that each different possible equilibrium price ratio would correspond to a different real allocation.

One very interesting case has been discussed recently. Suppose that we have some uncertainty about the future. There are no contingent markets for commodities; they can be purchased on spot markets after the uncertainty is resolved. However, there is a set of financial contingent securities, that is, insurance policies that pay off in money for each contingency. Purchasing power can therefore be reallocated across states of the world. If there are as many independent contingent securities as possible states of the world, the equilibrium is the same as the competitive equilibrium with complete markets, as already noted in Arrow (1953). Suppose there are fewer securities than states of the world. Then some recent and partly still unpublished literature (Duffie 1985; Werner 1985; Geanakoplos and Mas-Colell 1986) shows that the prices of the securities are arbitrary (the spot prices for commodities adjust accordingly). This is not just a numéraire problem; the corresponding set of equilibrium real allocation has a dimensionality equal to the number of states of nature.

A related model with a similar conclusion of a continuum of equilibria is the concept of 'sunspot' equilibria (Cass and Shell 1983). Suppose there is some uncertainty about an event that has in fact no impact on any of the data of the economy. Suppose there is a market for a complete set of commodity contracts contingent on the possible outcomes of the event, and later there are spot markets. However, some of those who will participate in the spot markets cannot participate in the contingent commodity markets, perhaps because they have not yet been born. Then there is a continuum of equilibria. One is indeed the equilibrium based on 'fundamentals,' in which the contingencies are ignored. But there are other equilibria that do depend on the contingency that becomes relevant merely because everyone believes it is relevant. The sunspot equilibria illustrate that Merton's insight was at least partially valid; we can have situations where social truth is essentially a matter of convention, not of underlying realities.

The Economic Role of Informational Differences

Let me mention briefly still another and counter-intuitive implication of thoroughgoing rationality. As I noted earlier, identical individuals do not trade. Models of the securities markets based on homogeneity of individuals would imply zero trade; all changes in information are reflected in price changes that just induce each trader to continue holding the same portfolio. It is a natural hypothesis that one cause of trading is difference of information. If I learn something that affects the price of a stock and others do not, it seems reasonable to postulate that I will have an opportunity to buy or sell it for profit.

A little thought reveals that, if the rationality of all parties is common knowledge, this cannot occur. A sale of existing securities is simply a complicated bet, that is, a zero-sum transaction (between individuals who are identical apart from information). If both are risk averters, they would certainly never bet or, more generally, buy or sell securities to each other if they had the same information. If they have different information, each one will consider that the other has some information that he or she does not possess. An offer to buy or sell itself conveys information. The offer itself says that the offerer is expecting an advantage to himself or herself and therefore a loss to the other party, at least as calculated on the offerer's information. If this analysis is somewhat refined, it is easy to see that no transaction will in fact take

place, though there will be some transfer of information as a result of the offer and rejection. The price will adjust to reflect the information of all parties, though not necessarily all the information.

Candidly, this outcome seems most unlikely. It leaves as explanation for trade in securities and commodity futures only the heterogeneity of the participants in matters other than information. However, the respects in which individuals differ change relatively slowly, and the large volume of rapid turnover can hardly be explained on this basis. More generally, the role of speculators and the volume of resources expended on informational services seem to require a subjective belief, at least, that buying and selling are based on changes in information.

Some Concluding Remarks

The main implication of this extensive examination of the use of the rationality concept in economic analysis is the extremely severe strain on information-gathering and computing abilities. Behaviour of this kind is incompatible with the limits of the human being, even augmented with artificial aids (which, so far, seem to have had a trivial effect on productivity and the efficiency of decision making). Obviously, I am accepting the insight of Herbert Simon (1957, chs 14, 15), on the importance of recognizing that rationality is bounded. I am simply trying to illustrate that many of the customary defences that economists use to argue, in effect, that decision problems are relatively simple break down as soon as market power and the incompleteness of markets are recognized.

But a few more lessons turned up. For one thing, the combination of rationality, incomplete markets, and equilibrium in many cases leads to very weak conclusions, in the sense that there are whole continua of equilibria. This, incidentally, is a conclusion that is being found increasingly in the analysis of games with structures extended over time; games are just another example of social interaction, so the common element is not surprising. The implications of this result are not clear. On the one hand, it may be that recognizing the limits on rationality will reduce the number of equilibria. On the other hand, the problem may lie in the concept of equilibrium.

Rationality also seems capable of leading to conclusions flatly contrary to observation. I have cited the implication that there can be no securities transactions due to differences of information. Other similar propositions can be advanced, including the well-known proposition that there cannot be any money lying in the street, because someone else would have picked it up already.

The next step in analysis, I would conjecture, is a more consistent assumption of computability in the formulation of economic hypotheses. This is likely to have its own difficulties because, of course, not everything is computable, and there will be in this sense an inherently unpredictable element in rational behaviour. Some will be glad of such a conclusion.

Reprinted from *Journal of Business*, 1986, vol. 59, no. 4, pt. 2.

See Also

- ▶ Models and Theory
- ▶ Preferences
- ▶ Rational Behaviour
- ▶ Rationality, Bounded

Bibliography

Allais, M. 1979. The so-called Allais paradox and rational decisions under uncertainty. In *Expected utility hypothesis and the Allais Paradox*, ed. M. Allais and O. Hagen. Boston: Reidel.

Arrow, K.J. 1953. Le rôle des valeurs boursières dans la répartition la meilleure des risques. In *Econométrie*. Paris: Centre National de la Recherche Scientifique.

Arrow, K.J. 1979. The property rights doctrine and demand revelation under incomplete information. In *Economics and human welfare*, ed. M.J. Boskin. New York: Academic Press.

Aumann, R.J. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1236–1239.

Cass, D., and K. Shell. 1983. Do sunspots matter? *Journal of Political Economy* 91: 193–227.

Chamberlin, E. 1950. *The theory of monopolistic competition*, 6th ed. Cambridge, MA: Harvard University Press.
Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.
Cournot, A.A. 1838. *Researches into the mathematical principles of the theory of wealth*. Trans. N.T. Bacon. New York: Macmillan, 1927.
Debreu, G. 1959. *Theory of value*. New York: Wiley.
Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–23.
Duffie, J.D. 1985. Stochastic equilibria with incomplete financial markets. Research Paper No. 811. Stanford: Stanford University, Graduate School of Business.
Geanakoplos, J., and A. Mas-Colell. 1986. Real indeterminacy with financial assets. Paper No. MSRI 717–86. Berkeley: Mathematical Science Research Institute.
Gossen, H.H. 1983. *The laws of human relations*. Cambridge, MA: MIT Press.
Grunberg, E., and F. Modigliani. 1954. The predictability of social events. *Journal of Political Economy* 62: 465–478.
Hicks, H.R. 1946. *Value and capital*, 2nd ed. Oxford: Clarendon.
Jevons, W.S. 1871. *The theory of political economy*, 5th edn; reprinted. New York: Kelley, 1965.
Knight, F. 1921. *Risk, uncertainty, and profit*. Boston: Houghton Mifflin.
Lewis, D. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Lucas, R., and T. Sargent. 1981. *Rational expectations and econometric practice,* 2 vols. Minneapolis: University of Minnesota Press.
Mantel, R. 1974. On the characterization of excess demand. *Journal of Economic Theory* 6: 345–354.
Marshall, A. 1920. *Principles of economics*, 8th edn; reprinted. New York: Macmillan, 1948.
Merton, R.K. 1957. The self-fulfilling prophecy. In *Social theory and social structure,* ed. R.K. Merton, revised and enlarged edn. Glencoe: Free Press.
Mill, J.S. 1848. *Principles of political economy*, 1909. London: Longmans, Green.
Morgenstern, O. 1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* 6: 337–357.
Riley, J.G. 1979. Informational equilibrium. *Econometica* 47: 331–360.
Shafer, W., and H. Sonnenschein. 1982. Market demand and excess demand functions. In *Handbook of mathematical economics*, 2nd ed, ed. K.J. Arrow and M. Intriligator. Amsterdam: North-Holland.
Simon, H. 1957. *Models of man*. New York: Wiley.
Sonnenschein, H. 1973. Do Walras's identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 345–354.
Spence, A.M. 1974. *Market signaling*. Cambridge, MA: Harvard University Press.
Stigler, G.J. 1961. The economics of information. *Journal of Political Economy* 69: 213–225.
Walras, L. 1874. *Elements of pure economics*. Trans. W. Jaffé. London: Allen & Unwin, 1954.
Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36: 110–119.
Working, H. 1953. Futures trading and hedging. *American Economic Review* 43: 314–343.

Economic Theory of the State

R. Jessop

The basic forms, social functions, institutional boundaries and legitimating principles of states vary across historical epochs and also differ among specific regimes in the same epoch. This makes it difficult (some would even say impossible) to develop a theory which applies to all states – whether in general or simply in their economic aspects. This entry limits itself to some economic aspects of the capitalist state.

The Capitalist Type of State

Capital accumulation has occurred under the most divergent state forms, but not all state forms are equally supportive of capital accumulation. Various attempts have been made to construct theoretically an ideal type of state which is both possible and particularly appropriate under capitalism without claiming, however, that this 'capitalist type of state' exists always and everywhere in capitalist societies. Among other characteristics of this state form, three institutional features are worth noting here: it has an effective monopoly of coercive power, its resources are purchased with money derived from taxation, and its activities are subject to the rule of law. Each of these features is not only compatible with but also potentially supportive of the capitalist economic order.

Firstly, the state is able to monopolize coercion because capital appropriates the surplus labour of

workers through the wage-relation rather than through extra-economic compulsion. This monopoly is also functional since it prevents particular economic agents from using direct force to subvert the free play of market forces. Secondly, state resources can be purchased because capitalism involves generalized commodity production and money mediates the exchange of all commodities (including labour-power). The state should raise monetary taxation because it cannot meet its reproduction needs by selling its own output, and cannot expropriate them forcibly only from those who happen to produce them without undermining the formal equality and property rights which underpin capitalism. Thirdly, the rule of law can exist because capitalism presupposes the formal freedom and equality of all economic agents. Only if it exists can such agents rely on a stable and impartial legal and political environment for their long-term economic activities.

These three institutional traits of the state facilitate capital accumulation. But they are neither logically nor historically necessary; nor, where they occur, do they guarantee accumulation. This is not simply because economic factors themselves engender recurrent crises within capitalism. There are also distinct political reasons. These are rooted in the institutional form of the state and in the struggles which occur around the nature and purposes of state power. To take only three examples. The institutional separation between the state and economy is crystallized above all in the state's legitimate coercive monopoly and its incarnation of national-popular unity vis-à-vis the antagonistic private interests of civil society. This means that the state has the political and ideological capacities to disturb as well as to promote capital accumulation. Nor does the tax form have any self-evident limits. It can produce fiscal crises and/or disproportions between state expenditure and the requirements of capital accumulation. Thirdly, because the rule of law implies formal neutrality towards particular economic agents, it is correspondingly inadequate as a steering mechanism. But more purposive, ad hoc, discretionary interventions can produce bureaucratic overload and also disrupt the labour process and capitalist market forces. Whether such problems occur depends not only on the form of the state and its integration into the circuit of capital but also on the changing balance of political forces.

Economic Aspects of the Capitalist State

Nowhere are economic systems self-reproducing, self-regulating and self-sufficient. They always depend on other institutional systems and the contingent support of non-economic forces. The capitalist state clearly has a key role in securing such institutional preconditions; and it is also the nodal site for political support. This does not mean, however, that one can enumerate a set of essential economic functions which must be performed by the capitalist state. Indeed, paraphrasing Max Weber's more general comment on the modern state, one could say: there are no economic activities which capitalist states have not at some time undertaken and none which they undertake invariably and exclusively. In particular the capitalist state is neither confined to producing 'public goods' nor is it the sole producer of such goods. Instead, even if certain broad developmental tendencies can be identified, its precise economic activities are always conjunctural. They are always influenced, furthermore, by political and ideological as well as economic factors.

Economic Periodization of the Capitalist State

The structural relations between state and economy and the forms of state intervention typically vary across time as well as nations. This has encouraged attempts at periodization. Although labels vary, four phases are often identified: mercantilism, liberal capitalism, (simple) monopoly capitalism and late (or state monopoly) capitalism. Without necessarily endorsing these attempts at periodization, the basic features of each stage can be presented as follows.

Under mercantilism state power is used to establish the dominance of the capital relation

and market forces. This is the period of primitive accumulation and capitalist manufacture and is associated historically with the absolutist state. Once this dominance is secured, a liberal phase is said to follow. This involves the nightwatchman state which is restricted to securing the general external conditions of production and has no significant directly economic role. The third phase is linked to the dominance of monopoly capital and the rise of imperialism. In this stage the state serves to regulate the economic dominance of monopoly capital, assumes an active role of managing the economic and political relations between organized capital and the labour movement, and also employs extra-economic coercion abroad in inter-imperialist competition. Next comes the state monopoly capitalist stage. State management of the domestic economy through taxation, state credit, public enterprise and/or the so-called military-industrial complex now have an increasingly important role; the welfare state system and collective consumption become central to the reproduction of labour-power and to political management; and international and transnational state organizations have a key role in managing the world economy. Not all national economies have experienced all four stages and much depends on the timing of their capitalist development and on their place within the international division of labour.

Such changes in the state's economic role also involve reorganizing its overall institutional form. Growing state intervention is typically associated with the strengthening of the executive at the expense of the legislative branch, the rise of functional (as opposed to territorial) representation closely tied to the administration, the increased importance of the state economic apparatus and the growing dominance of economic criteria within non-economic departments, and the decline of the substantive rule of law (as opposed to the simple maintenance of legal forms) in favour of more discretionary forms of intervention. Thus the growth of state economic intervention leaves neither the economy nor the state unchanged. The circuit of capitals is socialized through the state and the state is reorganized to reflect economic needs (cf. Poulantzas 1978).

Explanations for the Economic Role of the State

Various explanations have been offered for the state's assumption of economic functions and for their general developmental tendencies. Broadly speaking these comprise two main groups: explanations which focus on the essential structure and laws of motion of capitalism and explanations which focus on the social relations which obtain between class forces. Included among the former are explanations which emphasize the inability or failure of individual capitals, market forces or the law of value to secure all the institutional and economic conditions needed for capital accumulation. These conditions are frequently said to include: (a) 'general external conditions' such as bourgeois law or a formally rational monetary system; (b) public goods such as fire services, sea walls or statistical services which facilitate production in all branches; and (c) material factors productively consumed in all or most branches, such as labour-power or energy supplies. The more 'class-theoretical' explanations focus either on the state's instrumentalization by particular (capitalist) class interests and/or on its relatively autonomous role in managing the balance of class forces both within the economic sphere and in society more generally. In turn the relative strength of class forces is sometimes attributed to changes in the mode of production and sometimes to broader social and political factors ranging from unionization to wars.

The reasons advanced for increasing state intervention can be used to illustrate such arguments. Some theorists highlight changes in the forces of production (e.g their increasing socialization, growing capital intensity or lengthening turnover time of capital). Others emphasize changes in the relations of production (e.g. the shift from absolute to relative surplus value, growth of monopoly capital, increased importance of banking or financial capital, the internationalization of production, or changing forms of economic crisis). Yet others have stressed an increased importance of the tendency of the rate of profit to fall. Whatever reasons are advanced, however, the same conclusion is drawn. In the

course of capital accumulation there is a growing need for state intervention to socialize the forces of production (e.g. infrastructural provision, manpower training, technological innovation) and/or the relations of production (e.g. state credit, economic management or collective consumption) to compensate for the failures of market forces and competition adequately to coordinate and integrate the circuit of capital. Whilst such explanations often identify important structural changes in capitalism, they do not provide a satisfactory explanation for the political response to such changes. Nor does an emphasis on the mediating role of class struggle help much here unless attention is paid to the full range and forms of political forces.

The General Limits to State Intervention

There has been considerable interest in the limits as well as the reasons for state intervention. Again we find both general explanations and arguments relating to various stages of intervention. The following factors are frequently cited here: (a) the exclusion of the state from the heart of the production process – which means it must react *a posteriori* to events it cannot directly control or engage in ineffective *a priori* planning; (b) its tendency to respond to economic problems and crises in terms of surface appearances (e.g. inflation, unemployment, trade deficits) which have no obvious or consistent relationship to the real course of capital accumulation – which means that state policies often have limited or perverse effects; (c) the inherent limitations of law and money as steering mechanisms for a constitutional tax-state – since both mechanisms operate at a distance from real economic agents and processes; (d) the contradictions involved in the expansion of non-commodity forms of provision – they may promote capital accumulation but they also withdraw money from the circuit of capital, they can promote fiscal crises, and they suggest that the commodity form is neither natural nor necessary; and (e) the *sui generis* interests of state managers which can conflict with the supposed needs of capital. Most of these difficulties are aggravated by the co-existence of an effective world economy and a multiplicity of nation-states.

Political and Ideological Complications

The state's economic role is always affected by its other tasks. These include its own organizational reproduction, maintaining domestic political order and territorial integrity, and defining and interpreting national unity. Thus economic policies are typically inserted into more general political strategies and influenced by political and ideological struggles. This affects the inputs, 'withinputs' and outputs of the state system.

On the input side economic needs must be translated into political demands through whatever organizational and institutional channels are available; and they must be coupled with political values and legal norms which are often only indirectly relevant to economic considerations. Within the state system it is the balance of political forces which determines how these economic demands are expressed in economic policies. This will vary with the individual forms of policy production (e.g. bureaucratic, purposive programming, participation, delegation to professionals) and with the manner in which some basic unity is imposed on the state's manifold activities. Each mode of policy-production contains its own limitations; moreover, problems of internal unity often preclude the flexible responses needed for economic management. All this is aggravated because political forces are generally most immediately concerned with other political forces and only indirectly with the economic sphere. Accordingly, it is the political repercussions of economic events and crises which matter more than their inherent economic form or substance. Finally, the outputs of the state are generally mediated in and through its own forms of intervention which operate at one or more removes from the real economy.

Even the increasingly dominant state economic apparatus must operate in this environment and it is also prey to muddling through, administrative inertia, political pressures and ideological

thinking. State-owned industries and central banks typically operate in a political environment which shapes their economic activities and distinguish them from private industrial or financial enterprises. In general, state intervention reflects the balance among all political forces and these extend well beyond the classes, fractions and strata defined by the circuit of capital. This helps to explain the incoherence of economic policies and the difficulties of rational economic planning.

Indeed the state's current expanded role involves two double-binds: the one economic, the other political. Firstly, when the state intervenes to alleviate structural economic crises, it must substitute its own policies for the purgative effects of market-mediated reorganization. Thus it typically changes the forms in which economic crises operate rather than eliminating them and even internalizes such crises within the state. Here they can take such forms as fiscal crises, legitimacy crises, representational crises, crises of internal unity and crises of governmental effectiveness or overload. But, since the state's role has now become vital for accumulation, it cannot solve economic crises simply by withdrawing or refusing to intervene. At best it can reorganize how it intervenes. Moreover, in so far as economic crises are seen to follow from such withdrawal, refusal or reorganization, they can also precipitate new forms of political crisis. Secondly, in attempting to resolve crises on behalf of capital, it faces a political dilemma. If its crisis-management deliberately favours one fraction of capital at the expense of others, it is liable to aggravate economic problems for capital as a whole and to weaken its own legitimacy. But even if it succeeds in winning support for policies in the collective interests of capital, it cannot thereby avoid favouring some capitals more than others. This will modify the balance of forces and could disturb the initial alliance which sustained such policies.

Further Research

A general economic theory of the capitalist state is impossible because national economies and nation-states are too varied and because economic issues are always influenced by non-economic factors. But a theoretically informed account of the economic aspects of particular capitalist states is certainly possible. In this context it would be worth exploring the following issues. What forms are taken by the institutional separation of the state from the economic realm and what do these forms imply for the nature and limits of state intervention? How can one identify the collective interests of capital when these are always overdetermined by contingent political and ideological factors and when alternative paths and strategies are followed in different national economies? What difference do the various forms of political representation and intervention make to the economic role of the state in capitalist societies? What scope is there for international state organizations to regulate or manage economic crises? In answering such questions one must recognize that, despite the above-mentioned limitations to the state's capacities to manage capitalism, some states and regimes are more successful than others. This suggests the need for much more detailed historical analyses and for taking seriously the 'political' moment of political economy.

See Also

- ▶ Keynesianism
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Nationalization
- ▶ Welfare State

Bibliography

Alford, R.R., and R. Friedland. 1986. *Powers of theory: The state, capitalism, and democracy*. Cambridge: Cambridge University Press.

Badie, B., and P. Birnbaum. 1983. *The sociology of the state*. Chicago: Chicago University Press.

de Brunhoff, S. 1978. *The state, capital, and economic policy*. London: Pluto.

Galbraith, J.K. 1967. *The new industrial state*. London: André Deutsch.

Jessop, B. 1982. *The capitalist state*. Oxford/New York: Martin Robertson/New York University Press.

Kraetke, M. 1985. *Kritik der Finanzwissenschaft*. Frankfurt: VSA.

Luhmann, N. 1982. *Politische Theorie im Wohlfahrstaat*. Munich: Olzog.
O'Connor, J. 1973. *The fiscal crisis of the state*. London/New York: Macmillan/St Martin's.
Offe, C. 1984. *Contradictions of the welfare state*. London: Hutchinson.
Offe, C. 1985. *Disorganized capitalism*. Oxford: Polity Press.
Poggi, G. 1978. *The development of the modern state*. London/Stanford: Hutchinson/Stanford University Press.
Poulantzas, N. 1978. *State, power, socialism*. London: New Left Books.

Economic War

P. J. D. Wiles

Economic war constitutes all economic measures taken, before, during or instead of a military war, to harm an enemy. Compare protectionism, which is all the measures taken to 'defend' the national economy. These latter are often precisely the same measures. The subjective perception of how they do defend our own long-run economic interests is very often incorrect, and always controversial: for free trade lies at the root of Western economics. By contrast there is little theory about economic war, and (or so?) most of the measures taken seem by common admission well fitted to their time and place.

In view of the paucity of 'embargological' writing this entry must be of a frankly introductory character. First, it is well to establish some key definitions:

Embargo – a state's (or alliance's) prohibition to all its (their) citizens to sell to, or buy from, a named party, even when the price is right. An embargo is not an act of military war, and one on imports is little different from protectionism, except that its motive is to harm the foreign seller not benefit his domestic competitor.

Blockade – the prohibition by a state upon third states to trade with the second state, its enemy; a blockade must be enforced by military means and so is an act of war, possibly even against third states.

Both embargoes and blockades normally list specific goods and services. Note that the embargo of a sufficiently wide alliance is as good as a blockade, but is still no act of war.

Boycott – an embargo, usually popular or informal, on purchases alone. Typically the state machine is not involved, but some social group.

Contraband – goods on such a list that a third state tries to smuggle through a blockade.

Sanctions – the League of Nations' word for its members' punishment of an aggressor by a combined official blockade (the wording of Article 16 is vague in all original and amended versions, so the word 'blockade' is a little strong).

*Transport stranglehold*s – when one country's transport system monopolizes, or nearly so, access to another. The great case recently is Mozambique over Southern Rhodesia (the Beira railway, see below). A near case used to be the Arab League's use of the Suez Canal against Israel.

Black List – when the state imposing an embargo (or blockade) seeks to enforce it by a secondary embargo directly on specific firms within a third (capitalist) country, that are 'violating' the original embargo as they are of course entitled by international law and the law of their own state. The Arab League runs such a blacklist. The USA enforces its stricter view of CoCom in the same way. Communist enterprises are of course 'unblacklistable' – apart from their states.

Hostile Planner – the external authority who intervenes in the market (his, ours or the world's) in order to do us harm. This concept is necessary to remind us that interventions are not always benevolent. However, just as those of the friendly (and so mainly internal) planner may be mistaken and so maleficent, those of the hostile planner may be mistaken and so beneficient.

Bottleneck effect – when an unsubstitutable import is successfully embargoed, and some activity must, at least in the short run, be shut down.

CoCom – the Co-ordinating Committee of the NATO powers plus Japan, Australia and New Zealand. Administers an embargo of militarily significant industrial products. Is consultative only, each member remaining sovereign.

Dual Use – services like rail freight and goods like special steel and aero engines have dual military and civilian use. Thus an embargo on military goods *must* hit some civilian ones. Economic war has a very long history indeed. Its variety is best appreciated by considering its first in the Mercantilist era. In the 17th and 18th centuries the state was broadly proto-Keynesian. It sought to expand the quantity of money, in order to increase employment, encourage development and – above all – collect a gold stock in case of war. Economic war was the normal condition of Mercantilist international relations, interrupted only by military war. Although it certainly had military implications, it was not, as in the 20th century, a sign of extreme hostility or easily distinguished from mere protectionism.

But how does a bankless state acquire money? If, as was normal, it had no gold or silver mines and could not steal any in its colonial conquests it could only run a balance of payments surplus. This would not only bring in money, it would also set off the foreign-trade multiplier – a concept dimly perceived but not analysed; i.e. the new money would not be hoarded. So trade was a zero-sum game – the international division of labour dates only from Smith – and indeed a war of all against all in search of gold. Therefore one embargoed imports and encouraged exports, for one's own good. In peace time one did this *contra mundum*, in wartime one concentrated on one's enemy, doing oneself good and him harm all at once. Exporting to the enemy (except technology) was very patriotic, since it harmed him. Even military supplies were allowed (British cloth for the Grande Armée), though not actual arms. This was all an essentially monetary, not an input/output, view of economic war.

Banks and paper money added to but did not modify these policies, notably in the Napoleonic Wars. Paper money was regarded with extreme suspicion – a sign of national weakness even if convertible, since clearly the authorities had already failed to gather enough gold for all purposes. So we add to our goals the destruction of the enemy's convertibility. When Pitt went off gold it was a Napoleonic victory, due to France's superior exports (of wheat). The nature of this victory was that it was a blow to the morale of an enemy with a weak balance of payments. Drained externally of the means of internal payment, Pitt was faced with severe unemployment in Yorkshire, and a budget deficit if he wished to do something about it (in the absence of an existing and functioning welfare state). He therefore went off gold and printed the money – a defeat all in itself.

Let us jump to the 19th century, during which 'embargology' declined as free trade doctrine spread, and the notion spread that war is an epiphenomenon on the real, freely trading, peaceful, liberal, capitalist, democratic world of the planetary economy. In such an environment, where also in practice few wars were fought between major powers, there was no incentive even to consider economic measures short of war.

The 20th century has not forgotten the 19th, and it is only shamefacedly that it has reverted to the practice of the 17th and 18th centuries. Not accidentally protectionism has grown back too, but the two are not mixed up as under Mercantilism. All modern states have banks and paper money, and the monetary peculiarities of late Mercantilism have dropped away. With the welfare state and fiscal/monetary policy the modern state can sufficiently mitigate external crises to retain domestic political stability. Inconvertibility and inflation will not alter its warlike stance. Economic war has therefore – again very rationally – become an input–output matter, though the state of our enemy's gold reserve continues to be a preoccupation since gold is fungible into any input.

But the main change, surely due to 19th century example, is that economic war is no longer waged for economic ends (make him economically weaker so that I can be economically stronger, trade being a zero-sum game), but only for 'political' ends (make him economically weaker

so that I can be militarily stronger). We may even infer that civilization has advanced: dirtly tricks are no longer played by states merely for civilian gain. Let us examine a few examples of the new, more purely military economic war.

In the simplest case a specific export is embargoed to the enemy. If it is not a finished good, like a weapon, but an input (e.g. special steel) or both (e.g. refined petrol), our enemy must shut down some activity because of the bottleneck effect. This is the main weapon of modern economic war. If, however, his gold reserve is low and his balance of payments strained we may also embargo his exports, quite in general. This will force him to cut an import of his choice, and so suffer a mild bottleneck.

The practical complications are illuminating. Should the USA embargo the sale of wheat to the USSR, or should France embargo the purchase of Soviet gas? Provided that France does not become dependent on this gas (e.g. above five per cent of all fuel consumption) she clearly has a better economic case for doing what she prefers. For the USA, wheat relieves a serious and immediate bottleneck: that of fodder, leading to the immediate slaughter of Soviet livestock.

By a simple and well tried 'iteration', the livestock slaughter first raises, then lowers the supply of meat, the great crucial consumer good shortage that has already lead to very serious rioting and many deaths (Novocherkassk, 1962), not to mention a huge consumer subsidy. For comparison, in 1801 Britain imported French wheat to avoid a serious food shortage and despite Mercantilist doctrine. The mad Tsar Paul suggested a wheat embargo, this being his period of alliance with France. But that would have been to embargo an export, so everyone pointed out that he was only the mad Tsar Paul. Napoleon, of course, supported by current doctrine, had no qualms about his export. Anyway had not low farm prices contributed to the Vendée? Similarly Reagan fears, or feared, low farm prices, and brought Carter's wheat embargo to an end.

Yet again, wheat is a perfectly competitive commodity, and so much less suitable to be embargoed (though sometimes easy enough to blockade). In fact under President Carter the USSR bought wheat from Argentina instead. But the price was higher, the docking facilities worse and the delay considerable. All this imposed external costs the USSR, while USA, selling elsewhere in the world, had very minor external losses. Her losses were internal, indeed mainly only transfers, embarrassing the government but not much impoverishing the people: price support outlays, storage costs and electoral shifts.

Nevertheless it is part of the conventional wisdom of modern 'embargology' to count as far as possible in physical terms. The embargo deprived USSR of scarcely any bushels of wheat, so it is accounted a failure. The notion of a discriminatory export tax, of depressing the enemy's terms of trade, has achieved no recognition: the intellectual world of modern economic war is one of input-output and, seemingly, fixed co-efficients Mercantilism knew better. Even the export of money itself (long-term loans) is not taxed, but simply subjected to administrative control. But it has eventually been agreed, among the NATO powers, no longer to subsidize loans to Warsaw Pact countries; i.e. not to operate export credit guarantees in favour of even non-embargoed exports. At least, like machinery, large long-term loans are not perfectly competitive and so much easier to control.

If Mercantilism knew little about foreign lending, it knew as well as we do about technology transfer. Technology, like gold itself, was an exception: it must never be exported. For with better technology 'we' beat 'them', both in war and in the exportation of ordinary goods and services. In modern times technological levels differ much more, and the subject has become more important. Although no one country has a monopoly, the advanced have become very advanced, and it has become much more difficult to absorb their output; their active help is needed. There has also grown up an unduly sharp distinction between civilian and military technology – as if dual use were inconceivable. Moreover, military R&D bulks much larger in the total.

It was the beginning of the end of Mercantilism when David Hume declared that, In opposition to his narrow and malignant opinion, I will venture to assert, that the increase of riches and commerce

in any one nation, instead of hurting, commonly promote (sic) the riches and commerce of all its neighbours; and that a state can scarcely carry its trade and industry very far, where all the surrounding states are buried in ignorance, sloth and barbarism (*Three Essays*... II: *On the Jealousy of Trade*, Josiah Tucker's edn, London 1787, first page).

Economic war contributes much less than nothing if we only want to prosper. In the circumstances of the Cold War, the sole long-term economic war that the world now knows, this is clearly still true, but irrelevant. The great question is purely, will this new political system – opposed to 'us' on principle, and both expecting and working for 'our' total defeat – become more friendly just because it is richer? Or will it spend the extra resources on yet more arms?

The political aims of an economic war are seldom clear. Do we want (i) to incapacitate our enemy, (ii) to dissuade him, or (iii) much more ambitiously, to change his policy and aims? And with which economic instruments should we proceed in each case? In the absence of good theory modern political leaders enter upon economic war in permanent ignorance and temporary passion; their Mercantilist predecessors were far better served.

Case (iii) is bimodal. It includes, as a valid 'offensive' tactic, bringing the enemy into our group, transferring technology to him, lending him money at a discount and so enriching him: 'stab with a sausage'. In a basically economic analysis we need only say, this is absolutely correct, and the best policy by far, but only if it is sure to work, and within reasonable time. If not, case (iii) means, bimodally, that very severe measures indeed are appropriate: conversion through fear.

Case (ii) implies short slaps on the wrist, with valid threats of worse to come. It implies that we have *some* ability to change policy, at least in small matters, and are therefore prepared to 'fine-tune' our measures and to agree with each other on tactics.

Case (i) implies despair over ultimate friendship, and accepts a 'peace that is no peace' as a long-term goal: the establishment of military superiority by permanently slowing up the enemy's economic growth, without fine-tuning. One cannot after all fine-tune so diverse and fractious a coalition as the CoCom.

Modern economic war concerns mainly military and dual-use goods. This is an unnecessary restraint: if our enemy can make wheat with difficulty and rifles with ease we should deprive him of wheat. The logic is irrefragable in Case (iii) strategies, indeed hard to beat in Case (ii). Lipstick, therefore, is a highly strategic commodity if our enemy taxes it heavily and his comparative cost situation makes its production for any reason expensive for him. The concentration of embargoes in military goods serves however a good electoral purpose; ordinary people do not understand the lipstick argument but do agree that we should not deliver weapons (a not wholly correct proposition!).

Do the initiators of economic warfare always fail in their aims? This is often stated these days, by those who wish to end the CoCom and widen embargoes and with it (unilaterally) the Cold War against USSR. There is, however, no truth in 'always'; at most one can say, politicians initiate military war with far more thought, and it is not the fault of economic war, but of those who wage it, that its record is so spotty.

We list the main disputed or forgotten incidents since 1919:

1935–6. League of Nations sanctions against Italy, on the occasion of her Abyssinian aggression. Excessive moderation shown: neither oil imports nor use of the Suez Canal embargoed, but these were the only two serious bottlenecks. Reasons: fear of war in Mediterranean, and of Fascist-Nazi alliance.

1940. Anglo-American partial embargo on oil for Japan. Japanese general staff estimate military action will shortly become impossible. Pearl Habor results. This catastrophe for the initiators shows, at any rate, the effectiveness of their threat.

1976. Ian Smith, leader of the illegal white government of Southern Rhodesia, was forced to

go to the negotiating table with his black enemies by Samora Machel's closure of the Beira railway. In power since 1974, Machel had hesitated because of the huge loss of invisible earnings. The effect of this was to divert all traffic to the South African network, which is about five times as far to the sea, and so very expensive; overloading it was also very unpopular with the South African government (but to South African pressure was added greater guerrilla activity). So the success of the Mozambican embargo redeemed the failure of the British. The latter was of course grossly mis-conceived. Even if better administered it could not have worked before the Portuguese Revolution.

The beginning of East-West Détente in 1970 merits longer treatment. First Brezhnev offered the German Treaty, then the Helsinki Declaration and then, more informally, the emigration of Jews. These were, in their original form, substantial concessions, and the quid pro quo was to be technology transfer, and access to Western capital markets. In 1972 the deal was in place: the frontiers of West Berlin were recognized, the European Security Conference had begun (to end in 1975 with the Helsinki Declaration on human rights, communications, etc.), and the Jews were coming out.

But in the same year, 1972, Senator Jackson boasted during elections too much of how he had literally bargained the loans against the emigration. Sheer pride forced Brezhnev to hold back his emigrant Jews and the deal turned sour. This however does not alter the fact that the original détente was made possible by the US embargoes on technology and capital: the very Soviet political concessions basic to the earlier Détente, which the Western enemies of CoCom and the renewed Cold War wish to bring back, were themselves the product of the relaxation of the still earlier embargoes.

Economic war against South Africa since about 1946 has been, until September 1985, mainly a matter of private boycott; except that the Communist powers have embargoed her (save Mozambique, which is much more dependent than upon Southern Rhodesia; and the USSR which has co-operated in the international diamond duopoly). Ideologically motivated private groups in the advanced capitalist democracies have refused to buy this or that export; but since they have never fully controlled any enterprises this has affected only consumer goods. States have embargoed the sale of weapons and police equipment (except Israel and Brazil). All this is standard stuff – and was very ineffective.

Much more novel was the 'extra-territorial' use of shareholder power. Much as the US government forces its firms to boycott Swedish firms that have been blacklisted for ignoring CoCom, so have ideological groups of shareholders forced enterprises with branches in South Africa to raise black wages above the market level, recognize black unions and even to evade local laws. This has been achieved more by bad publicity than by serious voting blocks at shareholders' meetings. The role of the churches, both as shareholders and as propagandists, has been considerable. Such interference is known as extra-territoriality: the state (Sweden or South Africa) on whose territory the enterprise produces or sells, or the trade union organizes, loses the degree of control over events that is normal in a capitalist state owing to foreign bodies with their own political will. This is not the case if it has merely to deal with a profit seeking headquarters abroad. Non-profit seekers are much more formidable, once in full control.

Disinvestment runs clean contrary to this. Anti-Apartheid campaigners have divided into pragmatists wishing to use such little powers as extra-territoriality confers, and extremists wishing to keep, above all, their hands clean. Disinvestment is no weapon at all against a company that does not want to borrow more, and the refusal to recognize this simple fact shows us again at what a low intellectual level economic war is ordinarily discussed. But disinvestment has a corollary of very great potency indeed: the refusal to buy new issues. This refusal rubs off on the bonds and bills of the South African government. It

was of course the disinvestment controversy, and the spreading of the consciousness of what Apartheid really means, that made conservative Western banking circles refuse to 'put together a package' during the debt crisis of September 1985, turning them into a sort of moralized IMF. It will be observed that the more monetary, Mercantilist view of economic war has lost little validity.

Let us conclude with a mixed bag of applications of economic theory, for war and trade have many parallels:

(a) small countries are seldom in a position to make economic war, but are ideal victims of it:
(b) even large ones are not often well placed. Countries should form alliances, or coalitions as one says in oligopoly theory.
(c) even before size comes factor endowment. To be the monopolist of a raw material is great, but to possess an irreplaceable transport artery is still greater. And factor endowment is always largely historical chance.
(d) trade unions make economic war and throw up many parallels.
(e) to a most curious extent there is little notion of compensation for the losses caused to one's side by economic war. Once's image is of rich corporations losing small sums by not selling, or delaying the sale of, high technology. So the issue only arises domestically when small enterprises (e.g. farmers) are hit. As to international burden sharing, say with CoCom, the diplomacy of it would be horrendously complicated and divisive. But could not Britain have subsidized Mozambique, already in 1975, to close the Beira railway?

See Also

- ▶ Beggar-Thy-Neighbour
- ▶ Conflict and Settlement
- ▶ Dumping
- ▶ Free Trade and Protection
- ▶ International Trade
- ▶ Optimal Tariffs
- ▶ Tariffs

Economics in Belgium

Guido Erreygers

Abstract

There is no single Belgian school of economics, but a number of Belgian economists have made significant contributions, including Adolphe Quetelet, Ernest Solvay, Léon Dupriez, Paul Van Zeeland, Gaston Eyskens, Étienne Sadi Kirschen, Robert Triffin and Jacques Drèze.

Keywords

Belgium; De Man Plan; De Molinari, Gustave; Drèze, Jacques; European Economic Association; Free trade; Quetelet, Adolphe; Triffin, Robert

JEL Classifications

B00; B10; B20

Although no genuine Belgian school of economics has ever emerged, except perhaps in the second half of the 20th century, Belgian economists have made original and significant contributions to the discipline and played a major role in the creation of a European community of economists.

Before the First World War

When Belgium gained independence in 1830, economics as a scientific discipline virtually did not exist in the country. By the middle of the 19th century, however, most universities were offering courses on economic subjects, economists began forming associations, and international cooperation was actively pursued. Throughout the 19th century French economic thought was undoubtedly the main source of inspiration for economists working in Belgium, but British, German and Dutch economic schools were also influential. Another characteristic of that period was the

strong separation along ideological lines, with little interaction between Catholic economists, mainly associated to the Catholic University of Louvain, on one side, and liberal and socialist economists, associated to the Free University of Brussels and the State Universities of Ghent and Liège, on the other. The ideological divide is clearly visible in the rather biased account of the history of economic thought in 19th-century Belgium published by Michotte (1904).

Although not in the first place known as an economist, the polymath Adolphe Quetelet (1796–1874) needs to be mentioned for his path-breaking contributions with regard to the use of statistics in the social sciences. He introduced the notion of the 'average man', which in economics influenced the work of both the German Historical School and William Stanley Jevons (Mosselmans 2005). A fine example of pioneering statistical research is provided by the household surveys of Édouard Ducpétiaux (1804–1868), whose data were used by Ernst Engel to derive relationships between consumption and income.

Not surprisingly, many Belgian economists considered themselves to be part of the liberal family. The first generation of liberal economists is probably best represented by Charles De Brouckere (1796–1860), who combined careers in politics, academics and business. Together with Adolphe Le Hardy de Beaulieu (1814–1894), the Italian émigré Giovanni Arrivabene (1787–1881) and others, he founded a Belgian association of free-traders (Erreygers 2001). The main accomplishment of the association was the organization of the Congrès des Économistes in September 1847 in Brussels, the very first international conference of economists attended predominantly by ardent free-traders, and also by Karl Marx and Friedrich Engels. The next generation of liberal economists was headed by Gustave De Molinari (1819–1912), who advocated an extreme libertarian form of liberalism, opposing virtually any form of government intervention. Some consider him to have laid the foundations of free-market anarchism, also known as anarchocapitalism (Hart 1981–2). Although he spent much of his time in France, where he was for a long time editor of the *Journal des Économistes*, he played a very active role in Belgium. He founded and edited *L'Économiste Belge*, animated the Société Belge d'Économie Politique, and managed to breathe new life into the free-trade movement, both nationally and internationally (Van Dijck 2008). In many of these initiatives he found a fellow-traveller in Charles Le Hardy de Beaulieu (1816–1871), who published several textbooks on economics. It must be added, however, that few Belgian economists shared De Molinari's extreme view of liberalism.

In the second half of the 19th century Émile De Laveleye (1822–1892) was the country's most prominent economist. This prolific writer covering a wide area of topics had been strongly influenced by the French philosopher and Christian socialist François Huet, who taught at the University of Ghent. De Laveleye's economic publications included work on the origins and varieties of property rights, on bimetallism and on socialist doctrines. He was professor of political economy at the University of Liège, and authored an often reprinted textbook on economics. He considered his views to be close to those of John Stuart Mill and the German Historical School. Although very much appreciated by his contemporaries – he built up an impressive international network of colleagues and correspondents – his contributions lost most of their influence soon after he died.

At that time the wealthy industrialist Ernest Solvay (1838–1922), founder of the chemical firm Solvay & Cie, started to turn his attention to social and economic issues. He was convinced that a change of the monetary system (replacing the system based on metallic money by a pure-credit system which he called 'social comptabilism') combined with a sweeping reform of taxation (replacing all existing taxes by taxes on gifts and bequests, with rates increasing with the number of transfers) would provide the clue to solving society's problems (Erreygers 1998; Boianovsky and Erreygers 2005). Solvay, a prominent liberal, worked on these issues in close collaboration with leading socialist economists such as Hector Denis (1845–1913) and Émile Vandervelde (1866–1938). Their monetary

propositions led to a debate with Léon Walras, who saw a great similarity with his own views. On a more practical level, Solvay influenced economics in Belgium through his generous funding of various institutions associated to the University of Brussels, the most important of which are the Institut de Sociologie and the École de Commerce Solvay (now Solvay Business School). These institutions allowed economists such as Émile Waxweiler (1867–1916), Maurice Ansiaux (1869–1943) and Boris Chlepner (1890–1964) to do research.

Socialist doctrines found responsive audiences in Belgium, partly as the result of the rapid industrialization, but also because of the presence of exiles such as Karl Marx and Joseph Proudhon. The Saint-Simonians and Fourierists attracted scores of young intellectuals. This created a fertile ground for such figures as Hippolyte Colins de Ham (1783–1859), who proposed to provide all adults with a capital endowment, Joseph Charlier (1816–1896), who launched the idea of an unconditional basic income, and César De Paepe (1842–1890), who tried to bridge the gap between the Marxists and the anarchists (Cunliffe and Erreygers 2001).

At the University of Louvain economics had a decidedly Catholic profile in the 19th century. Charles Périn (1815–1905) and Victor Brants (1856–1917) both aimed at developing a social economics in accordance with the doctrine of the Church, along the lines of Le Play.

The Interwar Period

After the end of the First World War fellowships offered by the Commission for Relief in Belgium and the Belgian American Educational Foundation gave many talented economics students the opportunity to spend at least one year in the United States. This was the case with Léon Dupriez (1901–1986), Paul Van Zeeland (1893–1973) and Gaston Eyskens (1905–1988), who combined their studies at the University of Louvain with stays in respectively Harvard, Princeton and Columbia. As a result Belgian economics gradually obtained a more American character and the University of Louvain became much more prominent in economic research (Maes and Buyst 2005).

Dupriez introduced to Belgium the statistical business cycle techniques used by Harvard University. He became the driving force of the Institut des Sciences Économiques (later renamed Institut de Recherches Économiques et Sociales, IRES), founded in 1928 at the University of Louvain. Van Zeeland, who had joined the National Bank of Belgium as head of its research department, made a swift career in the bank and was soon considered as one of the country's leading economic experts. In the troubled political and economic climate of the 1930s he was twice prime minister of governments of 'national unity'. With the scientific backing of IRES, Van Zeeland successfully devalued the Belgian currency in 1936. The Van Zeeland governments also included the socialist Hendrik De Man (1885–1951), who in 1933 had proposed an ambitious 'Labour Plan' (also known as the 'De Man Plan') as a way out of the economic depression. His project, which was intensively discussed and had broad support in Belgium and in other countries, involved state planning of the economy and a technocratic way of governing.

A few isolated attempts were made to introduce a more mathematical approach in economics. The most ambitious project was that of Bernard Chait (1893–1957), an engineer and businessman who was in close contact with Jan Tinbergen and François Divisia. In his 1938 Ph.D. thesis he constructed a general mathematical theory capable of explaining business cycle movements, but he failed to convince economists of its usefulness (Erreygers and Jolink 2007).

After the Second World War

In the first half of the 20th century economics gained increasing recognition as a mature scientific discipline. Universities created special

schools and separate faculties for economic sciences, and several economics journals were launched. In the northern part of the country Dutch gradually replaced French as the language of instruction. Both the Universities of Brussels and Louvain were eventually split into Dutch-speaking and French-speaking universities.

As early as the 1930s Gaston Eyskens had become the leading figure of the Dutch-speaking section of economists at the University of Louvain. In the years after the war he seemed to be more receptive to the ideas of Keynes than the leading figure on the French-speaking side, Léon Dupriez, who favoured a laissez-faire approach and resisted to the introduction of Keynesian macroeconomic policies. The tension between the Dutch-speaking and French-speaking section at the university led to the foundation, in 1955, of a separate Dutch-speaking research institute at the University of Louvain, the Center for Economic Studies. The Center provided scientific backing for various economic reforms adopted under Eyskens's period as prime minister of Belgium (1958–61, 1968–72). Eyskens also took the initiative to create a government planning bureau (Buyst et al. 2005).

At the University of Brussels Étienne Sadi Kirschen (1913–2000) was the main agent of change in the 1950s. After having worked at the Office of European Economic Cooperation in Paris, he organized a team which estimated the first national accounts for Belgium. In 1957 Kirschen and his collaborators founded the Département d'Économie Appliquée, better known under its acronym DULBEA, the economic research institute of the University of Brussels. It constructed the first input–output table for Belgium. In close cooperation with Tinbergen, who lectured at the University of Brussels in the 1960s, Kirschen emphasized policy-oriented research based on quantitative methods, as exemplified in the ambitious threevolume *Economic Policy in our Time* (1964) by an international team of economists led by Kirschen (Sirjacobs 1997).

A few Belgian economists emigrated and made a career abroad. The most striking case is that of Robert Triffin (1911–1993), who initially studied at the University of Louvain. Having obtained his Ph.D. at Harvard, he worked for the Fed and the International Monetary Fund (IMF) – under its first director, the former Belgian Finance Minister Camille Gutt (1884–1971) – and in 1951 became professor of economics at Yale University. He made himself a reputation by pointing out a crucial weakness of the Bretton Woods system (later known as the Triffin dilemma) and arguing for a fundamental reform of the international monetary system. He was an influential economic adviser to key players in the European economic and monetary integration process; he returned to Belgium in the 1970s (Maes and Buyst 2005). A less well-known émigré is Raymond De Roover (1904–1972), who after his studies at the Antwerp Catholic business school decided to specialize in economic history. He earned his Ph.D. at the University of Chicago, and was later appointed as professor in Boston and New York. His work on early banking in Bruges and Florence made him a leading specialist on late medieval economic history and thought.

Probably the most important development in Belgian economics after the Second World War was initiated by a man who decided not to stay in the United States after completing his Ph.D. thesis. In 1966 the Center for Operations Research and Econometrics (CORE) was founded in Louvain. This was very much the achievement of Jacques Drèze (b. 1929), a student of the University of Liège who thanks to a fellowship of the Commission for Relief in Belgium went to the United States and obtained a Ph.D. at Columbia. After his return to Belgium he was appointed at the University of Louvain, where he rapidly replaced Dupriez as the dominant figure of the economics faculty, but he kept close contacts with his American colleagues, especially at Northwestern and Chicago, where he was visiting professor in the 1960s. The creation of CORE marked the adoption of a decidedly American style of doing research. From the outset CORE was meant as an interdisciplinary research institute, bringing together specialists in econometrics, statistics, operations research, game theory and

(mathematical) economics. Thanks to grants from the Ford Foundation and other institutions, it created a stimulating environment for research and offered fellowships to both Ph.D. students and established researchers. CORE was unique not only because it brought together Dutch-speaking and French-speaking economists from the University of Louvain, but also because Drèze managed to get the econometricians of the University of Brussels involved in the project. Moreover, from the very beginning CORE opened itself to the world: it hired the Dutch econometrician Anton Barten (b. 1930), established strong links with other European institutions focusing on quantitative economics, and welcomed American and other foreign scholars as fellows (Maes and Buyst 2005).

The University of Brussels and CORE played a major role in the creation of the *European Economic Review* and the European Economic Association. The *European Economic Review* was founded in 1969 by the European Scientific Association for Medium and Long Term Economic Forecasting (ASEPELT), of which Kirschen was the driving force. His younger colleagues Jean Waelbroeck (b. 1927) and Herbert Glejser (b. 1938), both of the University of Brussels, were the founding editors. In 1985 Waelbroeck and three other Belgian economists, Jean Jaskold Gabszewicz (b. 1936), Louis Phlips (b. 1933) and Jacques-François Thisse (b. 1946), all affiliated to CORE, took the initiative to launch the European Economic Association. Drèze was elected as the first president.

Drèze's contributions to economics are extensive, covering uncertainty, general equilibrium theory, macroeconomics, econometrics and much more (Dehez and Licandro 2005). He is by far the most influential Belgian economist of the second half of the 20th century. Besides those already mentioned, other important Belgian economists include Claude d'Aspremont (b. 1946), Paul De Grauwe (b. 1946), Mathias Dewatripont (b. 1959), Pierre Pestieau (b. 1943), Jean-Philippe Platteau (b. 1947) and Gérard Roland (b. 1954). It is remarkable that the top of the economics profession remains very much dominated by French-speaking economists.

See Also

- ▶ Catholic Economic Thought
- ▶ Tinbergen, Jan (1903–1994)
- ▶ Triffin, Robert (1911–1993)

Bibliography

Boianovsky, M., and G. Erreygers. 2005. Social comptabilism and pure credit systems: Solvay and Wicksell on monetary reform. In *The experiment in the history of economics*, ed. P. Fontaine and R. Leonard, 98–134. London: Routledge.

Buyst, E., I. Maes, H.W. Plasmeijer, and E. Schoorl. 2005. Comparing the development of economics during the twentieth century in Belgium and the Netherlands. *History of Political Economy* 37: 61–78.

Cunliffe, J., and G. Erreygers. 2001. The enigmatic legacy of Fourier: Joseph Charlier and basic income. *History of Political Economy* 33: 459–484.

Dehez, P., and O. Licandro. 2005. From uncertainty to macroeconomics and back: An interview with Jacques Drèze. *Macroeconomic Dynamics* 9: 429–461.

Erreygers, G. 1998. The economic theories and social reform proposals of Ernest Solvay (1838–1922). In *European economists of the early 20th century. Volume One: Studies of neglected thinkers of Belgium, France, The Netherlands and Scandinavia*, ed. W. Samuels, 220–262. Cheltenham: Edward Elgar.

Erreygers, G. 2001. Economic associations in Belgium in the 19th century. In *The spread of political economy and the professionalisation of economists: Economic societies in Europe, America and Japan in the nineteenth century*, ed. M. Augello and M. Guidi, 91–108. London: Routledge.

Erreygers, G., and A. Jolink. 2007. Perturbation, networks and business cycles: Bernard Chait's pioneering work in econometrics. *European Journal of the History of Economic Thought* 14: 543–571.

Hart, D.M. 1981–2. Gustave de Molinari and the anti-statist liberal tradition. *Journal of Libertarian Studies* 5: 263–290, 399–434; 6: 83–104.

Maes, I., and E. Buyst. 2005. Migration and Americanization: The special case of Belgian economics. *European Journal of the History of Economic Thought* 12: 73–88.

Michotte, P. 1904. *Études sur les theories économiques qui dominèrent en Belgique de 1830 à 1886*. Louvain: Charles Peeters.

Mosselmans, B. 2005. Adolphe Quetelet, the average man and the development of economic methodology. *European Journal of the History of Economic Thought* 12: 565–582.

Sirjacobs, I. 1997. *L'économiste dans le temps*. Bruxelles: Archives de l'ULB.

Van Dijck, M. 2008. From science to popularization, and back: The science and journalism of the Belgian economist Gustave de Molinari. *Science in Context* 21: 377–402.

Economics Libraries and Documentation

P. Sturges

Libraries for a discipline are formed and characterized by that discipline. It is quite easy for a visitor to recognize that a library is for scientists or for humanists or for social scientists just by a glance at the types of books, periodicals and other material on the shelves. Libraries for economists reflect the distinctive and changing sources and documentation of economics, which are in turn products of changes in the discipline itself. As economists have successively widened the scope of their enquiries and added weapons to their methodological armoury, so the types of material they have needed to consult have multiplied. As the immediate communication of their results has become more and more pressing, so the types of publication they have favoured have evolved in response. The library providing effective service to an econometrician today would have been as irrelevant to a 17th-century mercantilist as the literature of econometrics would have been incomprehensible. This article will deal with economics libraries in their natural context of economics documentation.

The first economics collections were in the private libraries of 17th- and 18th-century scholars. Adam Smith, renowned for his forgetfulness and carelessness in dress, was able to defend himself with the claim that at least 'I am a beau in my books'. In fact the majority of his 3000 titles were in the miscellaneous topics a gentleman scholar might have been expected to cultivate, with only about 100 directly on economic topics. Nevertheless, his economic method was a book-based one, using material from his predecessors and building it into his own system. About 100 authors are quoted in the *Wealth of Nations* (1776), though not always by name. By Smith's day there was already a large monograph literature the economist could draw on in several languages: pamphlets advocating trading schemes, tracts on farm management, alarming arguments on the growth or decline of population, accounts of particular industries, countries, cities, etc. The first economists wrote in monograph form too, either the pamphlet directed at some particular case or instance, or, from the 18th century onwards, the full length treatise summing up a whole theory of economics. Such was to remain the pattern of economics publishing until the later 19th century.

Not all economists relied on published sources; Malthus and others travelled widely in search of facts which they blended with information derived from extensive reading. There were other economists like Ricardo, whose theorizing worked outwards from his own inner store of business experience and personal insights, for whom access to a library was of less significance. An economist whose method did rely on books was usually forced to be a collector himself since libraries well-stocked with suitable publications were few. The 19th-century British economist, Richard Jones, in the laborious progress of his treatise on *Rent* (1831) made frequent calls on the kindness of his friend William Whewell to provide him with books from the 'Public Library' (in fact the predecessor of Cambridge University Library). As a country clergyman with little money for book buying and in comparative isolation from access to libraries, it was a difficult assignment for him to develop a theory of rent based on worldwide evidence. Indeed, as the complexity and specialization of the economics discipline grew during the 19th century it became less and less possible for an economist to function away from well-stocked libraries.

One reason for the increased dependence was the question of priority. As the literature of economics grew, so did the possibility of the prior publication of a supposed original idea. W.S. Jevons, for instance, developed and published his marginal utility theory in virtual isolation from other ideas on the topic. His subsequent realization that Menger and Walras had arrived at the same theoretical point in the same year of 1871 was disturbing, but as a voracious collector of economics literature he began to appreciate that there were also predecessors,

most notably Hermann Heinrich Gossen, whose work published in 1854, needed to be acknowledged. It was becoming clear in various inescapable ways that the economist needed a library. Great general libraries such as that of the British Museum, could and still can provide much for the economist. Karl Marx, for instance, laboured there to great effect for years, and was able to display an encyclopedic knowledge of the relevant literature as a result. Nevertheless, by the end of the 19th century, the discipline was ready for specialized economics libraries.

Estimates of the numbers of practising economists at a given time are even today fraught with problems of definition. However, it is fairly safe to say that until the end of the 19th century and the early 20th century brought the creation of schools and faculties of economics in the universities of Europe and North America, and governments and business began to hire people with the degrees awarded by these institutions, practising economists were very few in number and more significantly were thinly scattered geographically. A few, most notably Smith himself but others such as Jean-Baptiste Say, were professors in faculties of law or other marginally related disciplines. Most were first and foremost otherwise employed: as a businessman like Ricardo, an official like J.S. Mill, a clergyman like the abovementioned Richard Jones, an engineer like Dupuit, or a landowner like von Thünen. Libraries to serve such a scattered and heterogenous group were not practical. With the rise of Cambridge as a centre of economics excellence in the second half of the 19th century, the creation of the London School of Economics in 1895, and the emergence of great faculties of economics at Harvard, Columbia and Chicago by the beginning of the 20th century in the USA, there were for the first time concentrations of economics teachers, researchers and students, whose needs gave rise to specialized economics libraries.

The source materials that such libraries might stock were multiplying fast. The business world was the chief generator of publications of use to economists. Journals from many countries in specific fields such as mining, insurance and banking, prospectuses and annual reports of railway companies, histories of firms or industries, biographies of businessmen, documents from international fairs and exhibitions, all burgeoned during the 19th century. For the first time, governments became significant publishers of economics-related literature: British Parliamentary Papers and United States Congressional Documents began to appear with increased volume and regularity almost immediately the 19th century began. In other countries the quantity, if not the regularity, of government publications also increased swiftly as the century progressed. Much of their content was statistical; and the US Census of 1790 and the British Census of 1801 were milestones in the practice of number gathering. Statistical societies, such as that of London (founded 1825, now the Royal Statistical Society) arose to take advantage of this material and in the process created a new layer of publication which economists could exploit.

The growth of economics as a discipline not only brought about the multiplication of materials but also created a need for more immediate channels of communication. The monograph which had been the chief avenue of publication for so long, began to lose some of its importance to the periodical. During the 19th century, economists had written for the great literary reviews, for general interest magazines, for newspapers, and for the one or two specialized publications such as the *Economist*. It was only after 1886, however, when the *Quarterly Journal of Economics* was founded by Harvard University, followed quite swiftly by the Royal Economic Society's *Economic Journal* (1891) and Chicago's *Journal of Political Economy* (1892), that there was a serious rival medium for economic writings. When the *American Economic Review* joined them in 1911, a main core of prestigious journals was taking shape, and the nature of the communication of economic ideas was completely altered. The need to convince a publisher of the validity and interest of one's ideas, or alternatively to publish at one's own expense, was replaced by interaction with an editor or editorial board drawn from one's peers. The possibility of quite swift and direct communication of one's ideas, even on highly technical matters, was opened up. Economics publication

became more focused and more isolated from public debate.

This rich growth of sources permitted economists to build more and more elaborate structures of argument and evidence. W.S. Jevon's perilous but gallant attempts to link cyclical commercial fluctuations back through the business statistics to agricultural yields, to weather data, and finally to sunspot activity in a chain of causality, was just a particularly bold exploitation of the wealth that was becoming available. Even a collector of books as enthusiastic as Jevons could not reasonably hope to acquire materials of the number and type required for work of this kind. It is no coincidence that Jevons was an enthusiastic member of the Library Association (UK) and published papers on the topic of libraries. Whilst the 18th-century or early 19th-century economist urgently needed to be a collector of books, an economist of the late 19th and early 20th century like F.Y. Edgeworth needed to own hardly any books and could rely on libraries entirely.

The economics libraries which were created to cope with the growing need of economic researchers for specialized materials were, in the first place, historical collections designed to allow the reconstruction of past theory and the recovery of past knowledge. Two of the greatest have their roots in the personal collecting of one man. H.S. Foxwell (1849–1936) bridged the age of the great amateur collectors and that of specialised professional libraries. He amassed two great collections during a lifetime of acquisition carried on at a level of bibliomania. His first collection, sold in 1901 to the Goldsmiths Company, is now the basis of London University Library's Goldsmiths' Collection. His second collection, begun immediately the Goldsmiths sale had put his finances to rights, was sold in 1929 to Harvard Business School, which took possession on Foxwell's death to form the Kress Library. Kress is in turn now only part of the Baker Library of Harvard Business School.

Other great academic libraries also follow this pattern of a core historical collection alongside a fast-growing and fast-changing current collection. Columbia University, with its Seligman Collection, Johns Hopkins with the Hutzler Collection (assembled by Jacob Hollander) and the University of Illinois with its Hollander Collection, show a similar pattern. In Britain, the historical riches of Goldsmiths' are complemented within the London University system by the London School of Economics' British Library of Political and Economic Science, founded in 1896. Other parts of the world also have libraries of similar scope – Japan, for instance, where the Menger, Schumpeter and Burt Franklin collections of Hitotsubashi University provide a basis of old economics material to underpin its modern collections. With business history now an accepted discipline in graduate business schools in North America, the older titles in these libraries' core collections are attracting renewed attention from researchers.

Other types of economic research library have grown up in the 20th century to supplement the provisions of the academic libraries. Departments of Government and their associated agencies are the most common alternative source of economics materials. In the USA in particular, excellent libraries of this type are numerous; the US Department of Commerce Library, founded in 1913, for instance, has extensive collections including much statistical material, and agricultural economics material is one of the chief strengths of the National Library of Agriculture, originally established as the Department of Agriculture Library in 1862. The Federal Trade Commission, the Treasury, the Federal Reserve System Board of Governors, and other US Government agencies have fine economics libraries. The Library of Congress too, in its Social Science Reading Room, gives access to an enormous wealth of economics material. Other countries also have government economics libraries, the UK Board of Trade Library being one particularly fine example. International organizations also support economics collections; the International Labour Organisation (ILO) in Geneva, maintains an extremely large library, accessible via a computerized retrieval system. The Chambre de Commerce et d'Industrie Library in Paris is an example of a large, modern economics library supported by a trade association. The greatest economics collection in the public library sector is that of the New York Public Library, whose

Economic Division has since 1919 produced the weekly *Bulletin of the Public Affairs Information Service*, the foremost index to the world's English language literature in the related fields of economics, finance, business, labour and public affairs.

Only a comparatively few libraries have the resources to devote the necessary attention to the full potential range of material required by economists and the techniques that will permit its efficient acquisition and retrieval. The Library of the Instituts für Weltwirtschaft in Kiel, Federal Republic of Germany, regarded by many as the premier economics library in the world, has been able to do this consistently since its foundation in 1914. Its catalogue of persons gives access to material not only via authors of books, articles and chapters, editors of books, symposia and journals, writers of prefaces and introductions, but also to material whose subject is a person or persons connected with economics. Its title catalogue includes not merely books, but also annual reports, serials, newspapers, collections, and the corporate bodies, congresses and conferences which publish material. The subject catalogue has geographical entries in addition to conventional subject headings. Cross reference cards are hardly ever needed, for as many copies of the full entry card as are necessary are entered in the relevant places in the catalogue. The work of the library is largely in the hands of professional economists whose subject expertise ensures the accuracy of subject cataloguing. Because the provision of good catalogues is never enough to ensure that the user obtains documents that relate to his interests, Kiel like many other special libraries alerts users to pertinent new acquisitions. Mark Perlman (1973) has justly called its methods 'the acme of the traditional approach to economics literature retrieval'.

In addition to dealing with the complexities of the multiplying sources for economic research, libraries have had to come to terms with yet another shift in the forms of economic communication. The urgency to establish the priority of ideas or to publish research before it becomes obsolete has placed strains on book and journal publishing with which they have been unable to cope. The average lag between submission of an article and its publication in a journal has increased over the years, and the maximum wait may be in excess of two years. This is despite the introduction of submission fees intended to reduce the number of submissions to some journals, the growth in the number of specialized journals and the introduction of a journal, *Economics Letters*, specifically designed to ensure swift publication of material. Academic monograph publishing is currently under the severest financial strains, with spiralling costs leading to higher price per copy to the consumer, with consequently reduced numbers of sales driving the unit cost up still further. In their editorial decisions, therefore, publishers are putting increasing weight on the market value of proposed titles.

The solution to this problem has been the increasing use of semi-published forms usually referred to as working papers. This form of distribution for an individual paper, reproduced by some inexpensive method and circulated via the writer's own institution's mailing list, causes confusion and distress to some librarians. Some libraries only acquire working papers if they are free, many do not catalogue them, some bind them in series, others do not. Some libraries avoid them altogether because of the difficulties they cause, and in the conviction that anything worthwhile which appears as a working paper will eventually be published in more permanent form. This is a serious disservice to economic scholarship. Roy Harrod (1969) said 'Mimeographed essays issued in advance of publication, if any, by the research unit of one university to the professors of other universities all over the world have come to constitute the main matter for reading, at least among theoretical economists.' Indeed, some of the difficulty of knowing all the writings of an economist as distinguished as the Norwegian Ragnar Frisch stems from the fact that he published so frequently in working paper form. The collection of working papers at Warwick University Library. (UK), their published *Economics Working Papers Bibliography* and their microform service are a major contribution to this problem area.

In the last quarter of the 20th century, the forms of source material for the economist have begun to challenge the service capacities of traditional

librarianship. Most particularly, published statistics are no longer adequate for the purposes of many economists: they are out-of-date when published; they are in summary rather than comprehensive form; and they require recording in computerized form so that they can be arranged, shuffled and manoeuvred into revealing forms by the researcher. The electronic publication of statistics, making them available originally in computerized form, is growing rapidly. Increasingly, economic research depends on access to suitable computer hardware, availability of programmes which will perform the required tasks, and tapes of the data, rather than on books or other paper formats. To some extent this trend tends to exclude the library from the research process, but that is not necessarily always the case.

The issue of whether an economics library should merely confine itself to searching bibliographic databases on behalf of its economist clients or whether it should go further in identifying and making available numeric databases, has already begun to be explored, for instance in the Economics Library of the Ministry of Agriculture Fisheries and Food in the UK (O'Sullivan 1982). The Baker Library at Harvard has long been involved with the acquisition of computerized data bases, but now has professional staff members who are also actively involved in reviewing, publicising, and manipulating numerical databases. What is more, to add the exploitation of numerical databases, whether commercial, such as the many provided by Evans Economics, Inc. (EEI), or government created, such as the US Bureau of Labor Statistics' LABSTAT, or the various services of the Bureau of Economic Analysis, to the functions of a library which is already managed by computerized systems often does not seem like a major problem. The availability of statistics in published or database form, the relative costs, types of series available in the alternative forms, the compatibility of computerized data with the systems available to the economist, are undeniably difficult issues. However, librarians in their new roles as information specialists are proving themselves capable of lending invaluable assistance to their clients with this type of material.

The potential of computerized systems for the swift transmission of information is likely to be further exploited. Electronic mail, for instance, already permits the flexible exchange of messages, long or short, amongst individuals or groups, by users of computer systems linked by telephone lines. The potential of electronic mail for the almost instant communication of research findings amongst a group of interested experts, an 'invisible college', is obvious. The electronic journal, which is at present being developed at a number of centres, seems likely to be an answer to the problem of the chronic pressure on economics journals. The electronic journal exists originally as a database controlled by an editor. Authors send their 'manuscripts' to the editor on-line from wherever their computer is located, the text can be refereed, edited and amended on-line, and then the subscribers read the journal on-line, printing out text on demand. The journal can be altered daily if new material is received and old material can be relegated to storage files. Within a group of specialists, united by the necessary machinery and associated financial arrangements, accurate, easily modified information can be available in a form more swift and convenient than any previously used. This new communication medium need not necessarily circumvent specialized libraries, which in future might actually provide facilities for such operations, for instance in the archiving of material from the electronic journal.

With their computer expertise, their staffs of specialists, and their vested interest in the storage and dissemination of information, there are very obviously roles which libraries are developing in the management of electronic information systems. Though there seems little indication of the large-scale return to conventional library-based scholarship in economics prophesied by Harry Johnson (1977), a total disregard of the value of historic and rare books collections already existing should not be contemplated. No discipline can safely neglect its past and it is to be hoped that economics libraries will cherish their treasures for the use of the minority of scholars whose approach is more reflective and retrospective. Nonetheless, the cutting edge of economics librarianship in the 21st century seems likely to be

as different from that of the 20th or the 19th as will be its forms of documentation, and indeed the economics discipline which generates them.

Bibliography

Cole, A.H. 1957. *The historical development of economic and business literature*. Kress Library Publication no. 12. Boston: Baker Library, Harvard Business School.

Fletcher, J. 1972. A view of the literature of economics. *Journal of Documentation* 28(4): 283–295.

Harrod, R. 1969. How can economists communicate? *Times Literary Supplement*, 24 July, 805–806.

Johnson, H.G. 1977. Methodologies of economics. In *The organisation and retrieval of economic knowledge*, ed. M. Perlman, 496–509. London: Macmillan.

Kindleberger, C. 1977. The use of libraries by economists: A personal view. In *The organisation and retrieval of economic knowledge*, ed. M. Perlman. London: Macmillan.

Koch, J.E., and J.M. Pask. 1980. Working papers in academic business libraries. *College and Research Libraries* 41(6): 517–523.

O'Sullivan, S.D.A. 1982. Numeric and bibliographic databases for agricultural statistics: Conflict or co-operation? In *Sixth international online information meeting*. Oxford: Learned Information.

Perlman, M. 1972. Economic libraries and collections. In *Encyclopedia of library and information science*, vol. 7. New York: Marcel Dekker.

Perlman, M. 1973. Editor's comment [on Kiel Instituts für Weltwirtschaft Library]. *Journal of Economic Literature* 11(1): 56–58.

Ruokonen, K. 1981. BILD-integrated online system for economic and business literature. *Tidskrift für Dokumentation* 37(3): 62–67.

Yohe, G.W. 1980. Current publication lags in economics journals. *Journal of Economic Literature* 18(3): 1050–1055.

Economics of Franchises

Roger D. Blair and Jessica S. Haynes

Abstract

Franchising, which is common in many advanced economies, is a contractual form of vertical integration. This article examines the economic rationale for choosing franchising over vertical integration. It also examines the influence of the franchisor's ability to maximize its own profit.

Keywords

Distribution systems; Franchisee; Franchising; Franchisor; Royalties; Vertical integration

JEL Classifications

L14; L24; M55

Franchising is a contractual form of vertical integration. A manufacturer, for example, produces a product that must be distributed to consumers. The manufacturer can perform the distribution function itself through a chain of retail outlets it owns and operates. When a manufacturer both produces and distributes its product, the firm is vertically integrated by ownership. The manufacturer can then control the retail promotion, customer service, pricing, product availability, delivery and other relevant decisions at the distribution stage. It does this through internal managerial decisions designed to maximize the overall profit of the firm. But this is not the only way to organize the production and distribution of the firm's output. Instead of having a network of its own distributors, it can license independent firms to perform the distribution function. These licensees are termed franchisees, and the distribution system is called a franchise system. The manufacturer then engages in a contractual form of vertical integration. The franchise contract gives the franchisee the right to distribute the manufacturer's product, but also allows the manufacturer to control retail promotion, customer service, resale prices (minimum or maximum), and a host of other things that are important to the manufacturer, who is now a franchisor. But if franchising is just a contractual form of vertical integration, what is the rationale for franchising?

Rationale for Franchising

Brands provide information to consumers in a mobile society, and so reduce search costs.

When a consumer visits a branded outlet, there are supposed to be no surprises – pleasant or otherwise. Thus, local residents go to Roger's Ribs and Jessica's Java while visitors go to Sonny's Barbecue and Starbucks. The consumer's increased reliance on brand names has resulted in the development of chains of retail outlets. Chains benefit from lower costs as a result of economies of bulk purchasing and economies of scale in production, new product development, and promotion. In principle, all of the retail outlets could be corporately owned and managed, but many of the resulting chains are organized as franchises.

Franchising permits specialization that may lead to higher overall profits. The franchisor is the innovator who specializes in developing the brand, exploiting economies of scale in production and promotion, and negotiating with vendors on behalf of the chain. The franchisees bring their entrepreneurial spirit to their locations. They also contribute their knowledge of the local market. The result is a synergistic effect which increases potential profits (Caves and Murphy 1976). Franchising succeeds because it allows each party to do what it does best. Further, franchise contracts deliberately organize this relationship to give the franchisor and franchisee incentives to work in tandem to increase revenues.

Types of Franchise

Although the lines drawn are somewhat arbitrary, franchising can be divided into *traditional* franchising and *business format* franchising. Traditional franchising is used by a manufacturer to distribute its product through distributors (the franchisees) that are specifically licensed to do so. Traditional franchising is found in several industries: automobiles, beer, gasoline, ice cream, and soft-drink bottling to name a few. Business format franchising involves the use of the franchisor's brand, trademark and trade dress, and distinctive way of supplying goods and services to the consumer. In this case, the franchisor develops and promotes the concept while the franchisees implement the concept and carry out local production and distribution. The 'quick serve restaurant' sector is a good example. with familiar names such as the US-based chains McDonald's, Pizza Hut, Subway and Taco Bell. There are many other sectors in business format franchising including accounting services, automobile servicing, health and fitness, hotels and motels, and real estate.

Franchisor Compensation in Traditional Franchising

Traditional franchisors sell products to their franchisees which are resold to consumers. The franchisor earns its profit on these sales to franchisees. The extent to which franchising is a good substitute for vertical integration depends on two critical factors: the market structure in distribution and the relative efficiency of franchisees versus employee-managers. Assuming that franchisees are neither more nor less efficient than employee-managers, a manufacturer will earn the same profit whether it performs the distribution function itself or franchises the distribution. Its profit will be the same because the cost of performing the distribution function will be in the same in either case. If franchisees are more efficient than employee-managers, then distribution costs will be lower, retail demand will be higher, or both. This will improve the manufacturer's profit, due to increased sales to the franchisees.

If there is monopoly at the distribution stage, the market structure is one of successive monopoly. Assuming equal efficiency, the manufacturer will prefer vertical integration rather than franchising. Compared with the case of competitive distribution, successive monopoly results in double marginalization, which leads to lower output and higher price. The manufacturer's profits are reduced below the level that would result from vertical integration because double marginalization reduces the derived demand for the product.

These ill effects can be offset with other contractual provisions such as maximum resale price constraints, minimum quantity standards and price advertising (Blair and Esquibel 1996). If these are effective, they should eliminate the

exercise of downstream monopoly power and improve the manufacturer's profits.

Things are more complicated when franchisees are more efficient than employee-managers. The effect of the increased efficiency is to shift the derived demand to the right. This, of course, tends to improve the manufacturer's profits. The exercise of monopoly power by the franchise, however, tends to decrease the manufacturer's profits. The net effect cannot be determined on a priori grounds. Presumably, when a manufacturer expects the net effect to be positive, franchising is selected.

Franchisor Compensation in Business Format Franchising

Business format franchisors do not sell products to their franchisees for resale. Instead, they provide a concept, a brand, and a distinctive way of doing business. Business format franchisors have many ways of charging their franchisees for using the licence: initial franchise fees, sales revenue royalties, output royalties, rent and sales of necessary inputs, to name a few.

Nearly all franchisors include franchise fees in their contracts. These fees are structured as either initial or periodic lump-sum payments, with the initial form being far more prevalent. Blair and Lafontaine (2005, p. 61) found that 99.2 percent of all franchisors use initial franchise fees. In principle, both forms allow the franchisor to capture the maximum profit available under vertical integration, provided that the franchisee can obtain all inputs at competitive prices. To realize this level of profit, the franchisor sets the initial franchise fee or the present value of the periodic payments equal to the present value of the stream of future operating profits that will be generated by the franchisee. This, in turn, will allow the franchisee to earn only a competitive return on its investment.

As in traditional franchising, the efficiency of the franchisees relative to employee-managers contributes to the level of profits the franchisor can obtain. If franchisees are more efficient, this will increase the franchisor's profit because franchisees will bid up the franchise fee until it is equal to the stream of future operating profits.

Although theoretically feasible, franchise fees generally are not set at such a high level. This is predominantly a result of the uncertainty surrounding future market conditions, future interactions between the parties, as well as franchisees' wealth constraints. Often franchise fees just cover the start-up costs of opening a new franchise location.

Franchisors can also extract revenue by charging franchisees royalties, based on either a percentage of their sales revenue or a fixed fee on output sold. Sales revenue royalties are the more prominent of the two forms and the second most utilized charge by franchisors, next to initial franchise fees (Blair and Lafontaine 2005, p. 66).

As long as the franchisee can acquire all inputs at competitive prices and a competitive market exists among local franchisees, the franchisor can extract the optimum level of profits with royalties. In the case of sales revenue royalties, the franchisor achieves this outcome by setting the royalty rate equal to the ratio of the difference between the profit maximizing price and marginal cost to the profit maximizing price. This rate will result in a post-royalty price equal to the franchisees' marginal cost. The royalties collected will be precisely equal to the maximum profits that vertical integration would yield.

Again, franchisees' efficiency relative to employee-managers impacts these profits and the decision to franchise. If the franchisees are more efficient, then the franchisor will earn more profit due to the increase in sales revenue. If instead the employee-manager is more efficient, then the company will choose vertical integration.

If the franchisee has local monopoly power, however, the franchisor will be unable to attain the maximum level of profit through a sales royalty. The ability to exert local market power makes price endogenous for franchisees and allows them to factor sales revenue royalties into output decisions. Therefore, no sales revenue royalty will allow the franchisor to capture the amount of profit available through vertical integration. An equivalent analysis holds for output-based royalties (Blair and Kaserman 1980). Output-based royalties, however, make little sense in business

format franchises because there are often too many products to make tracking outputs feasible.

Another way in which franchisors obtain revenue is through input requirements. In many business format franchise systems, the franchisees are required to buy inputs from the franchisor. For example, a pizza franchisee may be required to buy pizza dough and sauces from the franchisor. If the franchisor charges its franchisees prices above the competitive level, these tying arrangements provide an alternative way of extracting the profit that vertical integration would provide (Blair and Kaserman 1978). Again, however, this result holds only when there is competition among the franchisees; otherwise, the higher input prices cause distortions that reduce overall profits.

Choosing to Franchise

The choice of franchising ultimately depends on whether it is more profitable than vertical integration. This, in turn, depends on the structure of the particular market and the efficiency of franchisees relative to employee-managers. The less competitive the downstream market structure, the less attractive franchising becomes. But the more efficient franchisees are, the more attractive franchising becomes. The choice of whether and how to franchise will therefore vary from chain to chain.

See Also

- ▶ Franchising
- ▶ Vertical Integration

Bibliography

Blair, R.D., and A. Esquibel. 1996. Maximum resale price restraints in franchising. *Antitrust Law Journal* 65: 157–180.

Blair, R.D., and D.L. Kaserman. 1978. Vertical integration, tying, and antitrust policy. *American Economic Review* 68: 397–402.

Blair, R.D., and D.L. Kaserman. 1980. Vertical control with variable proportions: Ownership integration and contractual equivalents. *Southern Economic Journal* 47: 1118–1128.

Blair, R.D., and D.L. Kaserman. 1983. *Law and economics of vertical integration and control*. Orlando: Academic Press.

Blair, R.D., and F. Lafontaine. 2005. *The economics of franchising*. New York: Cambridge University Press.

Caves, R.E., and W.F. Murphy. 1976. Franchising: Firms, markets, and intangible assets. *Southern Economic Journal* 42: 572–586.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Kaufmann, P.J., and F. Lafontaine. 1994. Cost of control: The source of economic rents for McDonald's franchises. *Journal of Law and Economics* 57: 417–454.

Rubin, P. 1976. The theory of the firm and the structure of the franchise contract. *Journal of Law and Economics* 21: 223–233.

Spengler, J.J. 1950. Vertical integration and antitrust policy. *Journal of Political Economy* 58: 347–352.

Economics of Online Recruitment

Catherine Thomas

Abstract

Online recruitment describes hiring workers who were initially selected via the Internet. The practice is now widespread, with a majority of US jobseekers undertaking some online job search activity. The online intermediaries that allow employers and employees to connect with each other, leading to online recruitment, vary in scope. They range from websites that provide information about workers or openings, thereby facilitating search ('job boards' or 'job search engines') to websites that both provide information and enable employers and employees to interact online during the hiring process ('online labour markets'). A subset of online labour markets also provides an infrastructure that allows for online management of the work process and payment systems.

Keywords

Adverse selection; Communications technology; Globalization; Hiring; Intermediation; Job boards; Labour; Moral hazard; Online

labour markets; Offshoring; Organizational economics; Outsourcing; Personnel economics; Search models; Search engines; Services industries; Trade frictions; Websites

JEL Classifications
J6; L86

This article describes the development of intermediaries that govern online recruitment and the communications technology that has shaped the course of that development. It presents data collected from various sources about the extent and growth of online recruitment. It then discusses some of the features of online recruitment that are of interest for general economic research. These features include online labour market search and matching processes; supply and demand in online labour markets; the implications of online hiring for remote work where output can be delivered electronically; and the increased availability of data about labour market transactions in online settings.

Job Boards and Job-Search Engines

The Internet allows employers and employees to communicate prior to an employment contract in a variety of ways. The common feature of online recruitment intermediaries is that they provide the institutional framework, or infrastructure, that permits this communication. The first main way in which they facilitate communication is to make employers aware of potential employees, and vice versa. In the early 1990s, along with the development of the Internet came organisations that allowed employers to post online descriptions of job vacancies and allowed prospective employers to search those listings. Websites that perform this function are commonly referred to as 'job boards'. Some sites also allow workers to post their résumés online to bring them to the attention of recruiters. While recruitment via these sites is initiated online, the jobs that are filled via this recruitment channel span many types of work and a broad range of sectors in the traditional offline economy.

Early job boards ranged from being quite generalist in the type of work posted to being very specialised. Large corporations were some of the first recruiters to adopt online technology in their hiring processes, purchasing technology from specialised software providers but performing the intermediation functions of managing applications in-house. As early as 2001, 90% of large US corporations were recruiting via the internet (Cappelli 2001), with the corporate homepage often providing a 'careers' link for prospective applicants. Large newspapers also put their classified section's job listings online, and non-profit organisations such as the US military began to operate online recruitment services. During the late 1990s, in tandem with the Internet boom, a handful of online job-search engines grew to prominence in this industry. These include Indeed, CareerBuilder, Monster.com, HotJobs and SimplyHired. Some of these job boards (such as LinkUp, Indeed and SimplyHired) are metasearch engines, which means that they aggregate job postings from other boards, an activity known as 'scraping' or 'wrapping'. While these large websites are broad in scope, there are also many industry-specific sites. One example is the IT industry site DICE.com.

Many Internet job-search engines also provide additional services that assist in employer and employee search. Most of the large sites are continually evolving, offering job-seekers and potential employers increasingly sophisticated tools. For example, Monster.com – now the world's largest recruitment website, with over 1.3 million resumes being posted globally each month (Monster.com 2012) – offers workers tools including skills analysis, cover letter and résumé-writing tips, and career management services such as employer comparisons and salary and benefits analysis. In 2012, Monster launched BeKnown, a professional networking application on Facebook, to allow employers to manage their online recruitment activities.

Online search dominates job-seeking activity in the USA and in other developed countries. Jansen et al. (2005) estimated that by 2005, more than 52 million Americans had performed online job searches. By 2010, a group of the largest

online recruiters had 51.1 million visitors per month among them (comScore 2010). In 2012, 99% of the Fortune 1000 firms had purchased services from or utilised job postings on Monster.com. The Monster Employment Index (MEI), a survey measure available on the Monster.com website, tracks changes in online recruitment activity over time and across different industry sectors in the USA, Canada, Europe and India, by gathering and summarising data from a large representative selection of career web sites and online job listings.

Online Labour Market Platforms

Intermediaries have also arisen to facilitate the online recruitment of workers whose output can be delivered electronically and they look very different from the search engines described above. When the work can be done remotely, the ability to recruit workers online has an even greater transformational effect on recruitment activities because, while traditional labour markets are segmented across geographical boundaries and by distance, online labour markets for remote work can match employers and employees across the world. Blinder and Krueger (2009) estimate that up to 25% of US jobs can be offshored, made possible, in part, by electronic product delivery.

Websites that assist employers in finding workers to deliver electronic output began to emerge in the late 1990s and early 2000s. While these websites function as platforms that facilitate search, they also provide infrastructure that allows employers to oversee the different stages of the employment process. Website functions include hiring, project/work management, payment management and API (application programming interface). These services reduce the trade frictions associated with hiring and managing a single remote worker, as well as the barriers to the coordination of multiple remote workers. Horton (2010) refers to these platform websites as 'Online Labour Markets' (OLMs) and describes them as having three common features: (1) Labour is exchanged for money within the platform. As such, there are two parties involved in each transaction, and these platforms differ from online tools that deliver content without payment, such as Wikipedia and Linux. (2) Output is delivered electronically. (3) The allocation of labour is determined and payments for services are conducted within the platform. Horton then divides the OLMs that have these three features into two categories: 'spot' OLMs and 'contest' OLMs.

Spot OLMs allow employers and employees to contract with each other regarding a specific piece of work, at a given price, or for a specified length of time. The work contracted ranges from data entry, to website optimisation, to discrete programming tasks. Large intermediaries in this category include oDesk.com, ELance, RentACoder and Guru. Amazon's Mechanical Turk (AMT) is an example of a spot OLM that offers a slightly different set of services, providing employers with the opportunity to hire a large temporary workforce online to perform wide-ranging tasks. This platform is often used both as an experimental site for research (where the human response is of interest in itself) and for commercial tasks in which humans are more effective than computers (for example, identifying objects in a photograph or transcribing audio recordings).

Data on the total number of employees and postings on each of these sites are relatively sparse. A 2009 report estimates that ten of the largest platforms had over 2.3 million registered workers at that time, although these workers are likely to overlap across sites (SmartSheet 2009). The types of tasks performed via spot online labour market platforms range from small, automated tasks with low pay, to simple projects, to relatively complex integrated tasks. Examples include: finding a set of email addresses or prices; writing a product review; designing a website or a presentation; programming software; or developing an algorithm. The tasks span services and solutions that are built on top of Micro Tasks platforms (such as often found on AMT) or technology-assisted relationship management on top of project management (as offered by oDesk).

The second category of online labour market platforms comprises 'contest' websites. The

employer–employee relationship on contest OLMs functions differently from that in spot markets, in that firms requiring certain electronic output post a competition for that output and invite submissions on the website. Workers who wish to enter the competition undertake the work, and buyers either select a winner or opt not to do so. Intermediaries of this type include CrowdSpring, 99Designs, GeniusRocket and LogoTournament.

Technological and organisational advances within these platforms have led to the globalisation of the labour market for this type of work. The employers demanding these types of labour services are often located in different countries from where the most efficient potential employees can be found, and employers can recruit large numbers of foreign workers with the required skills within these sites. As an illustration, in June 2010, over 60% of the revenues paid for online work in oDesk.com came from employers located in the USA, while workers located in the USA accounted for less than 15% of the revenues. The country accounting for the largest share of revenues earned was India, at almost 30%. A majority of transactions on this website span international borders, and employers regularly employ workers in different countries simultaneously.

In addition, the flexibility provided in these virtual workplaces allows for labour market participation among groups of domestic and foreign workers that previously had limited employment opportunities. The discrete nature of jobs, the variation in job duration, and the worker's ability to perform independent work, perhaps at irregular hours, make part-time online work possible for them. The online workforce includes large groups of students, for example.

Employer–Employee Matching and Wage Determination in Online Labour Markets

Because labour market outcomes are one key determinant of individuals' economic activity, economists devote much time to understanding how labour markets work (Rogerson et al. 2005). The growing prevalence of recruitment in online labour markets is, hence, of great interest to researchers because the longstanding and important questions about all labour markets are paralleled in online settings. In addition, the contrast between offline and online labour markets is, in itself, revealing about frictions that exist in all labour markets.

In particular, economists study labour market efficiency, motivated by several key empirical facts. One critical question is why it is that unemployment often exists at the same time as unfilled job vacancies. Another central question involves the way in which wage levels are set for individual workers in employment: why is it that similar workers earn persistently different wage rates? For example, Mortensen (2003) finds that observable characteristics typically explain no more than 30% of observed wage variation in traditional labour markets. These two facts each suggest that trade frictions in labour markets prevent these markets from clearing at an equilibrium wage rate that reflects the interaction of competitive demand and supply. As noted by Rogerson et al. (2005, p. 960), the empirical evidence suggests that 'there is simply no such thing as a centralised market where buyers and sellers of labour meet and trade at a single price, as assumed in classical equilibrium theory'.

Labour market economists have shown that introducing employer and employee search to labour market models can generate theoretical predictions that mirror observed outcomes in these markets. Much of this research has its origins in work done by the winners of the 2010 Nobel Prize in economic science, Peter Diamond, Dale Mortensen and Christopher Pissarides. These models typically consider a forward-looking individual's job-search problem – the decision of whether to look for a job based on expected future wages and the likelihood of finding a job, as well as job-search costs and any benefits from remaining unemployed. Similarly, an employer's problem about whether to post a job and search for a worker can be modelled as a function of the probability of finding a worker, the value of finding a worker to accept the job, the

expected wage that it would cost the employer, and any search costs that would be incurred. Petrongolo and Pissarides (2001) survey early matching work in labour markets, and discuss how various models differ in the wage determination process, matching function, and in the other assumptions made.

Online recruitment is generally thought to lower the costs associated with job search, in comparison to offline labour markets. When employers can browse the résumés of hundreds or thousands of potential hires that have all been screened as possessing relevant qualifications and experience for the job, it is much easier to become informed about candidates. In addition, many intermediaries lower the costs of contacting the workers, and some even facilitate employer–-employee contracting. This can all be done in front of computers from remote locations. Most models of search and matching predict that lower search costs increase match efficiency and, depending on the structure of the model, often also increase the productivity of the resulting employment matches. Both these effects increase economic output and have the potential to transform a whole range of labour market outcomes, including wages, job duration, aggregate unemployment and aggregate productivity.

Online recruitment, especially in online labour market platforms, is also changing the nature of the bargaining game between employers and workers that determines workers' wages. In traditional models of labour markets, wages are often modelled as the outcome of Nash Bargaining over the surplus generated by the transaction, relative to each party's outside option and bargaining power. One implication of the increased ease of search and matching in online markets could well be that the outside options for employers and employees, as well as their relative bargaining powers, are different than in offline markets. It could be that a given employer, when searching online rather than offline, has a choice between larger numbers of potential hires that are closer substitutes and who have fewer valuable outside options. For a given worker, on the other hand, the ability to search online may increase his or her access to potential employment opportunities. The overall consequences for the wage determination process are likely to depend on the nature of the job posting and the set of skills required in the job, as well as the matching function in the online labour market in question.

Online Labour Supply and Demand

While lower search costs in online recruitment are predicted to increase the efficiency of labour market outcomes for a given set of employers and employees, it is also likely that lower costs will affect the number and type of workers applying for jobs. As the costs of applying for jobs fall, the number of applicants for any one job is predicted to increase, and employers may find that having many more applicants makes it harder and more costly to find a good fit among the candidates. In particular, online applicant profiles tend to contain lots of information described by Autor (2001) as 'low bandwidth', meaning that the information is not helpful to employers in distinguishing between candidates. In online markets, employers have less of the 'high bandwidth' information that can be more easily gathered during inperson meetings and that allows employers to determine the quality of the match for each potential candidate.

Perhaps more importantly, in addition to increasing the number of applicants, when employers solicit applications, the composition of workers applying for jobs may change. When it is costly to apply, workers with private information about their own suitability for the job will apply only if the probability of their being hired, which is increasing in their fit, is sufficiently high. When application costs fall, candidates who judge themselves to be less well suited to the job will choose to apply, increasing the problems associated with adverse selection (Akerlof 1970). Autor (2001) forecasts that this adverse selection of lower-quality – or worse-fit – applicants into the set of candidates for any given job could work to offset efficiency gains from lower search costs. There is some empirical evidence that is consistent with the presence of adverse selection in several online markets. Among the largest OLM platforms, a large share of posted jobs are never

filled despite attracting large numbers of applications, suggesting that distinguishing between candidates remains a costly undertaking for recruiters in these markets.

However, low search costs and more competition among workers (leading to lower wages) are also likely to bring employers into the market and affect the number and type of jobs that employers post. A 2009 report by Smartsheet.com comments that many buyers of online labour services are individual employers or small companies with one-off projects and limited in-house resources. Often, the work done online via online recruitment would have been infeasible to staff through traditional labour markets. This includes tasks such as image tagging, transcription and other large volume tasks that can be completed online at relatively low cost.

Additional Features of Interest in Online Recruitment for Electronically Delivered Work

In markets where adverse selection threatens to lead to market unravelling, efficiency can be increased through the development of costly signalling mechanisms (Spence 1973). In online labour markets, it is possible that good-quality workers might find it worthwhile to agree to work at a very low wage until their type is revealed. High future wages would compensate for the cost of sending this signal only for those workers who are subsequently revealed, on the job, to be high quality. Employers, then, would be able to distinguish good-quality from bad-quality workers by offering low initial wages. In practice, however, there is some evidence that current online labour market institutions are unable to facilitate the costly signalling required to separate good from bad workers. While workers with no established reputation often bid to work at very low hourly wages, there is still unemployment among these workers and, at the same time, job postings go unfilled. One potential explanation is that hiring online workers incurs costs for employers in addition to the wages paid in any transaction, and the wages paid to workers cannot fall to levels that are low enough to induce employers to hire a worker of unknown quality. For instance, workers typically cannot bid to work for negative wages on OLMs. Some online labour market platforms are developing alternative mechanisms that allow the quality of inexperienced workers to be credibly revealed. For example, Stanton and Thomas (2012) show that inexperienced workers affiliated with the small organisations in oDesk.com called outsourcing agencies are high-quality workers. Affiliation is associated with greater employment success on the site, which suggests that affiliation acts as a credible signal of quality.

When work is undertaken remotely by workers recruited online, employers might be particularly concerned about moral hazard. For example, when the work undertaken requires a worker to have access to proprietary systems and data, an employer takes on the risk that this information will be expropriated. There are also the more standard concerns about the possibility of worker shirking on the job. The largest online labour markets have been able to develop an institutional context that discourages worker moral hazard or, at least, insures employers against it. Most OLMs operate dispute resolution systems – tending to favour the employer – and monitoring systems that reduce contractual incompleteness. Many have wellfunctioning employer feedback systems that allow workers to build valuable reputations in the marketplace and increase incentives for good performance on the job.

The low observed wages, especially in online labour market transactions between employers in developed economies and workers in low-wage countries, have led to public debate about whether these marketplaces constitute 'digital sweatshops' (Zittrain 2009). There is concern that firms hire remote workers in online markets to engage in cross-border institutional arbitrage, allowing employers to avoid their local labour market regulations. It is, however, not clear that this view is held by the workers on these sites. In his 2011 paper, Horton surveys workers on AMT and finds that they perceive online employers to be slightly fairer and more honest than offline employers.

One final issue that arises as a consequence of the growth in online recruitment for work that can be delivered electronically is that the need for physical concentration of the workforce is greatly reduced. A distributed workforce that spans the confines of a traditional workplace may also render traditional organisational hierarchies redundant. Some managerial roles that involve coordinating teams and processing information may be replaced by online communications technology. By facilitating online remote work, it is likely that developments in online recruitment have consequences for how firms are structured and for the way in which economic activity is undertaken, in general, in the global economy.

Data Availability About Market Transactions

Overall, then, online recruitment opens up new channels of communication that can be characterised as reducing search costs and affecting the bargaining power of each party, perhaps making labour supply more competitive. In labour economics, both the matching function and the bargaining function are central in determining labour market outcomes (and, thus, individual and economy-wide economic outcomes). But, in addition to changing the nature of these functions, online recruitment offers unprecedented access to micro-level data about the recruitment process that will allow the nature of these functions to be studied in much greater detail. Precisely because interactions happen electronically, at arm's length, data on the nature of these interactions can be stored and retrieved by researchers *ex post*.

As an example, it is now possible to determine the profiles that employers choose to view, how long they look at each profile, and whom they choose to contact. In OLMs, researchers can observe entire employment histories for workers and entire hiring histories for employers. In some cases, these websites collect data on the bargaining process over wages, along with both employers' and employees' outside options, as revealed by other job offers or potential hires. One remaining drawback is that firms and workers cannot generally be easily tracked across different markets. With access to all these data, it becomes feasible to observe the impact of policy experiments at a level of detail that is not practical in offline labour markets. Nonetheless, the extent to which any inferences drawn are externally valid, for example, in offline labour markets, would depend, however, on the experiment in question. Researchers should bear in mind that the type of jobs (and the type of workers) on online sites are a selected sample from the overall labour market.

See Also

- ▶ Adverse Selection
- ▶ Electronic Commerce
- ▶ Globalization and Labour
- ▶ International Outsourcing
- ▶ Internet, Economics of the
- ▶ Labour Market Search
- ▶ Labour Markets
- ▶ Moral Hazard
- ▶ Technical Change

Bibliography

Akerlof, G. 1970. The market for 'lemons', quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84(3): 488–500.

Autor, D. 2001. Wiring the labor market. *Journal of Economic Perspectives* 15(1): 25–40.

Blinder, A., and A. Krueger. 2009. Alternative measures of offshorability: A survey approach. NBER Working Papers 15287, National Bureau of Economic Research, Inc.

Cappelli, P. 2001. Making the most of on-line recruiting. *Harvard Business Review*, March.

comScore. 2010. *comScore media metrix ranks top-growing properties and site categories for January 2010*. Available at: http://www.comscore.com/Press_Events/Press_Releases/2010/2/comScore_Media_Metrix_Ranks_Top-Growing_Properties_and_Site_Categories_for_January_2010. Accessed 15 Oct 2012.

Horton, J. 2010. Online labor markets. Working paper.

Horton, J. 2011. The condition of the Turking class: Are online employers fair and honest? *Economics Letters* 111(1): 10–12.

Jansen, B., K. Jansen, and A. Spink. 2005. Using the web to look for work. Implications for online job seeking and recruiting. *Internet Research* 15(1): 49–66.

Monster.com. 2012. Monster Internal Data, monthly average Q3.
Mortensen, D. 2003. *Wage dispersion*. Cambridge, MA: MIT Press.
Petrongolo, B., and C. Pissarides. 2001. Looking into the black box: A survey of the matching function. *Journal of Economic Literature* 39(2): 390–431.
Rogerson, R., R. Shimer, and R. Wright. 2005. Search-theoretic models of the labor market: A survey. *Journal of Economic Literature* 43(4): 959–988.
Smartsheet.com. 2009. *Paid crowdsourcing. Current state and progress toward mainstream business use*. Available at: http://www.smartsheet.com/files/haymaker/Paid%20Crowdsourcing%20Sept%202009%20-%20Release%20Version%20-%20Smartsheet.pdf. Accessed 15 Oct 2012.
Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87(3): 355–374.
Stanton, C., and C. Thomas. 2012. Landing the first job: The value of intermediaries in online hiring. Working paper.
Zittrain, J. 2009. Work the new digital sweatshops. *Newsweek*, December. Available at: http://www.thedailybeast.com/newsweek/2009/12/07/work-the-new-digital-sweatshops.html

Economics, Definition of

Roger E. Backhouse and Steven Medema

Abstract

Economics is difficult to define unambiguously, many definitions having been proposed as the subject has evolved. Definitions are *ex post* constructions, even rationalizations, but they can nonetheless influence what economist do and how they set about doing it. This article considers the main definitions from the late 18th century to the present, pointing out some of the ways in which changing views reflect and have influenced changes in the subject.

Keywords

Cameralism; Economic science; Economics imperialism; Economics, definition of; Institutional design; Marshall, As; Mill, J. S; Operations research; Political economy; Rational choice; Robbins, L. C; Samuelson, P. A; Scarcity; Steuart, Sir J; Value judgements

JEL Classifications

B0

The definition of economics has evolved significantly over time, influenced by and influencing the focus of economic study. The definition often attributed to Jacob Viner, 'economics is what economists do', reflects the difficulty of providing an unambiguous definition. The problem, of course, is that definitions of the field are proposed *ex post* in an attempt to impose order upon a body of work that has grown up as economists have sought to tackle diverse practical and intellectual problems. Viner's statement suggests that there is no need for a tight, specific definition of the subject, which may explain the tendency of economists blithely to ignore definitions, and hence to not analyse them in detail, except sporadically. However, definitions of the subject do have effects through influencing what economists choose to study and the methods they think legitimate for analysing them.

The root of the word 'economics' lies in the Greek *οίκονομία*, meaning the management of a household, as in Xenophon's *Οίκονομικος*, written around 400 BC. In the 18th century, the idea of efficiently providing for the wants of a household was extended to the nation as a whole, under the heading 'political economy', the term first used for the discipline that later became economics. The first systematic English-language book on the subject was James Steuart's *An Inquiry into the Principles of Political Oeconomy* (1767, p. 16). Though Steuart made an analogy between 'providing for all the wants of a family, with prudence and frugality' and doing the same for the state, there was a difference, for the ruler of the state could not direct people in the way that the head of a household was able to do. This had the consequence that,

> The great art therefore of political oeconomy is, first to adapt the different operations of it [the state] to the spirit, manners, habits and customs of the

> people; and afterwards to model these circumstances so, as to be able to introduce a set of new and more useful institutions. (Steuart 1767, p. 16)

No doubt influenced by German Cameralism, Steuart saw institutional design as lying at the heart of political economy. This usage was followed by Adam Smith, who saw political economy as 'a branch of the science of a statesman or legislator' with two objects: providing the people with 'plentiful revenue or subsistence' and providing the state with enough revenue to provide public services (Smith 1776, p. 428).

Many of the classical economists, however, disagreed with the focus on policy, arguing that political economy was concerned with the laws that govern the production, distribution and consumption of wealth, the clearest example of this being Jean Baptiste Say, whose major work (1803) is *Traité d'économie politique, ou simple exposition de la manière dont se forment, se distribuent et se consomment les richesses* (A treatise on political economy, or a simple account of the way in which wealth is formed, distributed and consumed). This definition formed the basis for Nassau Senior's *Outline of the Science of Political Economy* (1836) in which he argued that the science was based on four propositions, the first and most important of which was 'That every man desires to obtain additional Wealth with as little sacrifice as possible' (Senior 1836, p. 26).

Neither of these definitions was acceptable to John Stuart Mill, whose 'On the Definition of Political Economy; and the Method of Investigation Proper to It', first published in 1836, was the last of his *Essays on Some Unsettled Questions of Political Economy* (1844). To define political economy as the rules for making a nation rich was to confuse 'art' and 'science'. However, it was not enough to define it as the laws relating to the production and use of wealth, for these included many physical laws that lay outside its remit. He thus favoured a more limited definition: 'The science which treats of the production and distribution of wealth, so far as they depend upon the laws of human nature' or 'The science relating to the moral or psychological laws of the production and distribution of wealth' (Mill 1844, p. 318). Mill went on to argue that even this definition was too broad, for political economy related only to man in society.

The most significant challenge to this definition of political economy as, loosely, the science of wealth, came from Alfred Marshall, who offered the well-known definition:

> Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing. (Marshall 1890, p. 1)

This definition is significant not so much for changing the name of the discipline to economics as for its focus on the study of mankind. For Marshall, as for many of his generation, the evolution of human character was of crucial importance: it was important to study actual human behaviour, but it was important, especially in the longer run, to consider how activities and consumption served to influence character and hence behaviour. Wants could not be taken as given but depended on activities.

In these discussions there was, as Neville Keynes pointed out, an ambiguity in the use of the word 'economic'. On the one hand it referred to attaining an end 'with the least possible expenditure of money, time and effort' (Keynes 1891, pp. 1–2) whilst on the other hand it was used as an adjective corresponding to the noun, wealth. The economists who laid most emphasis on the first of these were the Austrians – Carl Menger and his successors – who focused on economizing behaviour. It was his familiarity with this literature that led Lionel Robbins to deny originality for his much-quoted definition, 'Economics is the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses' (Robbins 1932, p. 16). Robbins's definition put scarcity and choice at the centre of economic analysis. He emphasized that 'any kind of human behaviour' that demonstrates the scarcity aspect falls within the scope of economics, and that there are 'no limitations on the subject-matter of Economic Science' beyond involving 'the relinquishment of other desired alternatives' (choice) (1932, p. 17).

The significance of this definition lies in its analytical nature: instead of defining economics in terms of its subject matter, it defines it as an aspect of behaviour. In spite of Robbins's claim that he was simply describing professional practice, the initial reaction of the profession to his definition of economics, at least as it surfaced in academic journal articles and introductory textbooks (where the definition of economics was primarily discussed), was negative (for a detailed discussion, see Backhouse and Medema 2007). Throughout the 1930s and 1940s, textbook writers continued to define economics in terms more reminiscent of Mill and Marshall than Robbins, in that, even where reference was made to scarcity, this was frequently qualified: economics was described as a social science concerned with the study of wealth, of earning a living or a study of the system of free enterprise. Robbins's choice-based definition was seen as too wide, and needed to be restricted so as to rule out matters that did not come within the 'traditional' boundaries of economics. The acceptance of the Robbins definition came piecemeal. First, scarcity came to be stressed as important to the subject. The first edition of Paul Samuelson's *Economics* (1948), undoubtedly the leading textbook in the post-war period, captures well the qualified attitude with which the Robbins definition was approached. Samuelson explained that economics was about scarcity, for 'the American way of life' required more resources than were available, but he chose to define the subject in terms of 'what', 'how' and 'for whom' – that is, as concerning the production and consumption of goods and services. There is nothing here that is inconsistent with Robbins, but this approach was equally consistent with a more traditional approach. Books such as George Stigler's *Theory of Price* (1946), which adopted the Robbins definition, laid great stress on both scarcity and choice, but others carefully refrained from doing so.

It was only in the late 1950s and 1960s that the use of Robbins's definition became widespread. By the late 1960s, Samuelson's *Economics* was claiming that economists agreed on 'a general definition something like the following':

> Economics is the study of how men and society *choose*, with or without the use of money, to employ *scarce* productive resources, which could have alternative uses, to produce various commodities over time and distribute them for consumption, now and in the future, among various people and groups in society. (Samuelson 1967, p. 5).

However, support for this was still not universal. For Richard Lipsey, whose *Introduction to Positive Economics* was one of the most successful rivals to Samuelson's *Economics*, scarcity was '*one of* the basic problems encountered in *most* aspects of economics', not the entire subject (Lipsey 1963/71, p. 50). Economics also dealt with questions related to failure to achieve a point on the production possibility frontier, such as explaining unemployment, which could not be reduced to problems of scarcity.

The move by Robbins to define economics as an aspect of behaviour made it just a short step to defining economics in terms of a method – that of rational choice-which could be applied not simply to production and consumption choices, but to all of human behaviour. This move was encouraged by the tendency, in the aftermath of the Second World War, to see economics though the lens of operations research, as social engineering, in which optimization techniques were central and game theory played a significant role. It has also been argued that this move towards emphasising rational choice had ideological attractions during the Cold War. During the 1960s, economics became increasingly conceived as the 'science of choice', without reference to a particular social domain, even, at times, without reference to scarcity: the subject could encompass non-market as well as market activities. The work of Theodore Schulz and Gary Becker on human capital, James Buchanan, Anthony Downs and Gordon Tullock on political processes, and Becker on discrimination and on crime and punishment laid foundation for what came to be called 'economics imperialism', the application of economics to fields including politics, law, history, and sociology. These theoretical moves were reinforced by advances on the empirical side, where the techniques developed by, for example, James Heckman and Daniel McFadden for analysing

cross-section data sets on individuals and households were used to investigate phenomena, such as non-marital fertility, that lie outside the traditional domain of economics as concerned with market behaviour.

Robbins's definition of economics in terms of the allocation of scarce resources remains the most widely cited definition of the subject, but it has never commanded universal assent. Though scarcity can be defined in such a way as to make it true, there have always been significant numbers of economists who have considered that it does not encompass all aspects of their discipline and that qualifications or extensions are required. These result in definitions closer to those found in the 19th-century literature, focusing on phenomena such as the production and distribution of wealth. At the other end of the spectrum, there are economists for whom rational choice is more fundamental than scarcity. To this extent, then, there is no universally agreed upon definition of the subject.

The reason this does not present a problem is that economists can proceed with their work irrespective of how their subject is defined. Definitions of fields generally come only after the field is established; as fields change, so definitions change. Despite this, however, definitions can matter. As Mill recognized, questions of method and definition are linked. The clearest example of this is Robbins, who sought to derive all the main propositions of economics from the premise of scarcity. His definition, therefore, was the basis for claiming that economic theory was central to economics – that it was far more important than Marshall had believed it to be. Also significant was his reference to economic science, for the word science is far from neutral. Robbins had argued that value judgements, including those necessary to make interpersonal welfare comparisons, did not come within the scope of economic science, but belonged instead to the realm of 'political economy'. In claiming this, he was arguably attempting to clarify the status of economists' arguments, for, as he later made very clear, offering any advice on economic policy requires such value judgements. Thus if economics includes policy advice it must encompass more than economic science as Robbins defines it. However, such is the prestige of 'science' that Robbins's definition caused many economists to try to dispense with value judgements altogether, even in welfare economics. An exercise in clarification (and no doubt a critique of certain views of the subject) thus had the effect of significantly narrowing the subject. Attempting to define economics thus was not and is not simply a descriptive exercise; it has consequences for what economists do, and how they go about doing it.

See Also

- ▶ Altruism, history of the Concept
- ▶ Mill, John Stuart (1806–1873)
- ▶ Rationality, History of the Concept
- ▶ Robbins, Lionel Charles (1898–1984)
- ▶ Samuelson, Paul Anthony (1915–2009)
- ▶ United States, Economics in (1945 to present)

Bibliography

Backhouse, R.E., and, S.G. Medema. 2007. Defining economics: Robbins's definition in theory and practice. SSRN working paper. Abstract online. Available at http://ssrn.com/abstract=969994. Accessed 19 May 2007.

Keynes, J.N. 1891. *The scope and method of political economy*. London: Macmillan.

Lipsey, R.G. 1963. *An introduction to positive economics*, 3rd ed. London: Weidenfeld & Nicolson, 1971.

Marshall, A. 1890. *Principles of economics*. London: Macmillan, 1949.

Mill, J.S. 1844. Essays on some unsettled questions of political economy. In *Collected works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press, 1967.

Robbins, L.C. 1932. *An essay on the nature and significance of economic science*, 2nd ed. London: Macmillan, 1935.

Samuelson, P.A. 1948. *Economics*. New York: McGraw Hill.

Samuelson, P.A. 1967. *Economics*, 7th ed. New York: McGraw Hill.

Say, J.-B. 1803. *Traité d'économie politique, ou simple exposition de la manière dont se forment, se distribuent et se consomment les richesses*. Paris: Deterville.

Senior, N. 1836. *An outline of the science of political economy*. New York: Augustus Kelley, 1965.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 2 vols. Indianapolis: Liberty Press, 1976.
Steuart, J. 1767. *An inquiry into the principles of political oeconomy*, 2 vols. Edinburgh: Oliver & Boyd, 1966.
Stigler, G.J. 1946. *The theory of price*. New York: Macmillan.

Economies and Diseconomies of Scale

Joaquim Silvestre

Conceptual Issues

Definitions

We consider the unit costs of producing a (single or composite) output under a given technology (no technical change). We say that there are *economies* (or *diseconomies*) *of scale* in some interval of output if the average cost is decreasing (or increasing) there. This definition focuses on economies and diseconomies of a technical character. It is sometimes extended to cover business activities other than production (such as marketing, financing, training: see Scherer 1980).

Note that, in the case of a composite output, the proportions among the goods produced are kept constant. (A different notion, that of 'economies of scope' contemplates variations in cost as the output mix varies.) The definition of cost may, on the other hand, imply that the input proportions are adjusted in order to minimize expenditures. A related idea is that of returns to scale: here both the output and input proportions are kept fixed, and one compares the amount of (the simple or composite) output $f(x)$ produced by a given input vector x with the amount produced by vector λx, for $\lambda > 1$. *Increasing* (or *decreasing*, or *constant*) *returns to scale* are said to prevail if $f(\lambda x)$ is greater than (or smaller than, or equal to) $\lambda f(x)$. Under some conditions (see, e.g., Fuss and McFadden 1978, p. 48) increasing (or decreasing) returns to scale are equivalent to economies (or diseconomies) of scale.

If f is a strictly concave function and $f(0) \geq 0$, or if f is homogeneous of degree less than one, then decreasing returns to scale prevail. Conversely, homogeneity of degree greater than one is a sufficient condition for increasing returns to scale.

Internal and External Economies and Diseconomies

It is sometimes useful (see, e.g., section, "Perfect Competition as Price Taking Behaviour" below) to consider economies of scale that appear only at the aggregate level and not at the level of the individual firm. For example (see Chipman 1970) let there be two firms with cost functions $C_j(y_j) = k_j y_j$, $j = 1, 2$. Firm j treats k_j as a parameter, and in this sense its technology displays constant returns to scale. But suppose that k_j actually depends on the amount of output of the other firm, say $k_j = [y_i]^{\beta}$. Then the aggregate cost is $[y_2]^{\beta} y_1 + [y_1]^{\beta} y_2$. We have *external economies* if $\beta < 0$ and *external diseconomies* if $\beta > 0$.

Explaining Diseconomies and Economies of Scale

We consider diseconomies first. Decreasing returns imply that duplicating *all* inputs yields less than twice the amount of output. But an exact clone of a production process that exhaustively lists all factors of production should give exactly the same output. The failure to double the output suggests the presence of an extra input, not listed among the arguments in the production function, that cannot be duplicated. This idea goes back to Ricardo's rent as based on the impossibility of duplicating agricultural land of a given quality. Alternatively, the extra input can be interpreted as managerial skill.

Consider (see McKenzie 1959) a strictly concave production function $f(x)$, where x is an L-dimensional input vector. One can associate to it a constant returns to scale technology with $L + 1$ inputs $F(x; z)$ and a fixed level of the extra input, say $z = 1$, such that $f(x) = F(x; 1)$, i.e., f describes the amounts of output obtainable by varying the first L inputs when the 'managerial skill' is kept at the constant level $z = 1$. To this end, define $F(x; z) = zf(x/z)$. It is easy to check that F is quasiconcave and homogeneous of

degree one, i.e., constant returns to scale. Moreover, competitive profits can be viewed as the competitive reward to the 'managerial skill' (at $z = 1$, $z(\partial F/\partial z) = f(x) - \nabla f(x) \cdot x$).

A similar notion can be applied to the case of external diseconomies of scale: the extra input can then be identified with a common pool resource (say, clean water), available in a limited amount. Conversely, the extra public input may be created by the activity of the industry (say, information or specific training of the labour pool): this will generate external economies of scale.

We turn now to internal economies of scale. Koopmans (1957) reviews some controversies on this issue and remarks (p. 152 fn.3), 'I have not found one example of increasing returns to scale where there is not some indivisible commodity in the surrounding circumstances.' The following ideas have appeared in the literature.

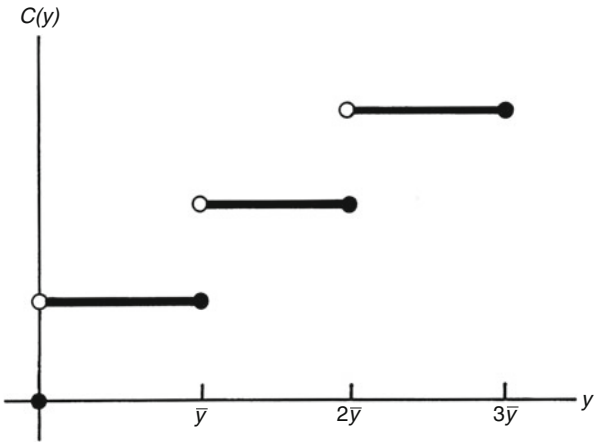
(a) *Indivisible input.* Assume for instance that the only input is some specific capital good (a machine, plant, ship or pipeline) which is indivisible in the sense that it becomes useless if physically divided. It has a given maximal capacity $\bar{y}$, but it can be underutilized to produce amounts of output less than $\bar{y}$. Then $C(y)$ looks like Fig. 1, and there are economies of scale in each of the intervals $[0, \bar{y}\,]$, $[\bar{y}, 2\bar{y}\,]$,K, $[(n-1)\bar{y}, n\bar{y}\,]$, K.

(b) *Set-up cost.* Take the only input to be labour time and assume that a certain amount of time has to be spent in preparation for the task (the set-up cost can be given several interpretations, as time spent in: (1) concentrating and getting psychologically ready for the task; (2) learning how to do it; (3) preparing the tools needed). Once the set-up cost is paid, the amount of output is proportional to the extra labour spent. This looks like Fig. 2, where increasing returns to scale prevail. Set-up costs can here be viewed as a form of indivisibility: 'readiness' (or 'information' or 'preparation') is indivisible: a 'half-ready' worker is useless.

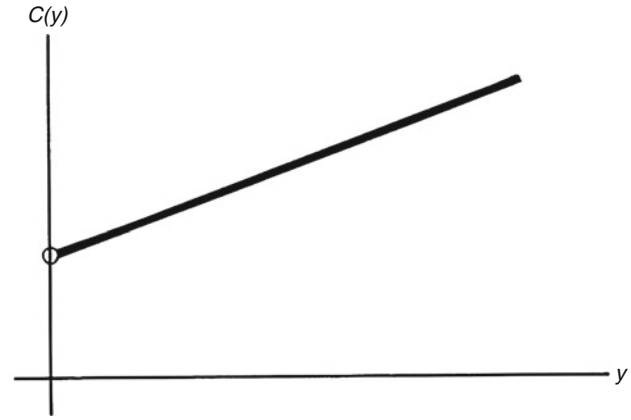
(c) The above examples can be extended to more than one capital good (or type of set-up cost). Consider, for instance, pipelines ten miles long. Only metal sheet is used in their production: the amount needed is proportional to the radius of the pipeline. Output y (flow of oil between two points ten miles apart) is proportional to the section area, a quadratic function of the radius. A pipeline of a given radius is indivisible, but one can build pipelines of any radius. The cost function looks like Fig. 3. The vertical coordinate can be interpreted as the minimal dollar outlay of a firm that buys pipelines and sells y, or as the amount of the input 'square yards of metal sheet' used by a vertically integrated firm that produces its own pipelines and sells output y.

(d) Adam Smith's division of labour. The *Wealth of Nations* attributes to the 'division of labour' the increase in output per worker. The main argument seems to be based on the set-up costs of (b) above. Smith's notion is related

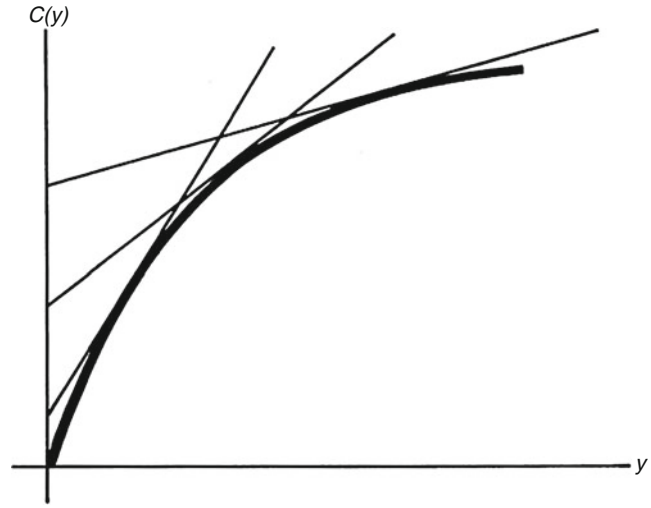
Economies and Diseconomies of Scale, Fig. 1



Economies and Diseconomies of Scale, Fig. 2



Economies and Diseconomies of Scale, Fig. 3



to another fundamental idea: the Ricardian gains from specialization and trade. But, in Arrow's (1979) words, 'the Ricardian idea of specialization lacks some characteristics of the Smithian; in Ricardo's system the abilities to produce are given. In Smith's view, specialization is more a matter of deliberate choice.'

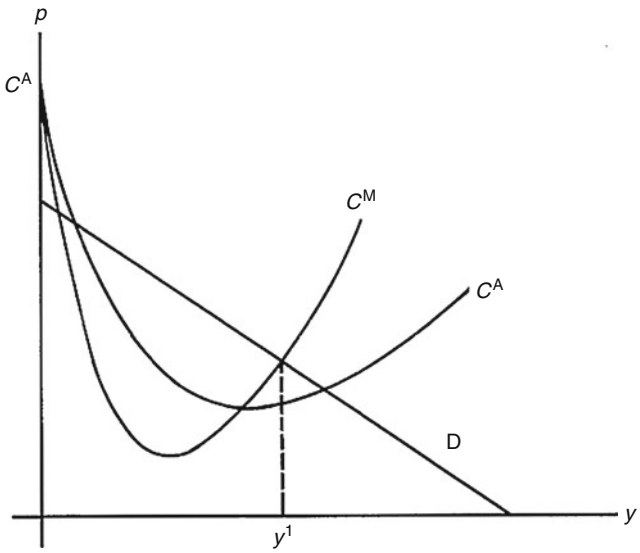
Economies of Scale and Market Structure

Perfect Competition as Price Taking Behaviour

A perfectly competitive firm is often defined as one that faces a horizontal demand curve. It is clear that, as long as $C(0) = 0$, no such firm can be at equilibrium at a level of output at which the average cost faced by the firm is decreasing.

This in particular implies that competition cannot prevail under the presence of internal economies of scale at all levels of output. The argument allows for economies of scale which are external to the competitive firm. (But the *laissez-faire* equilibrium will then typically be suboptimal.) The idea of constant (or increasing) average cost at the firm level but of decreasing average cost at the aggregate level (industry economies of scale) did play an important role in the controversies on the compatibility between economies of scale and competition (see Chipman 1965). This idea is illustrated in the example of section, "Internal and External Economies and Diseconomies"

Economies and Diseconomies of Scale, Fig. 4



above: the reader is referred to Chipman (1970) for a rigorous study.

We now focus on internal economies. Let the market demand curve be as in Fig. 4 where the average and marginal curves of a typical firm are also drawn. This situation does not *per se* violate the price taking rule: one could have, for instance, a single price taking firm operating at y^1. But such a combination of demand and cost does not fit well with the idea of perfect competition.

First, it is graphically clear that at most two firms may operate in this market. But it is then unrealistic to assume that each firm will take the market price (or the price charged by the other firm) as given.

Second, with two firms the aggregate supply curve would look like the discontinuous curve in Fig. 5, where supply at no price equals demand. Difficulties with the existence of competitive equilibrium will in general appear as soon as the average cost is somewhere decreasing.

These difficulties become more severe when considering unrestricted entry (with identical cost curves for all incumbents and entrants), a natural attribute of perfect competition: it is then required that at a 'long run' equilibrium no potential entrant have incentives to enter. But if potential entrants are themselves price takers, a long-run equilibrium implies zero profits, and typically none will exist if the cost curves are U-shaped. The entry model of Baumol et al. (1982) faces the same existence difficulties.

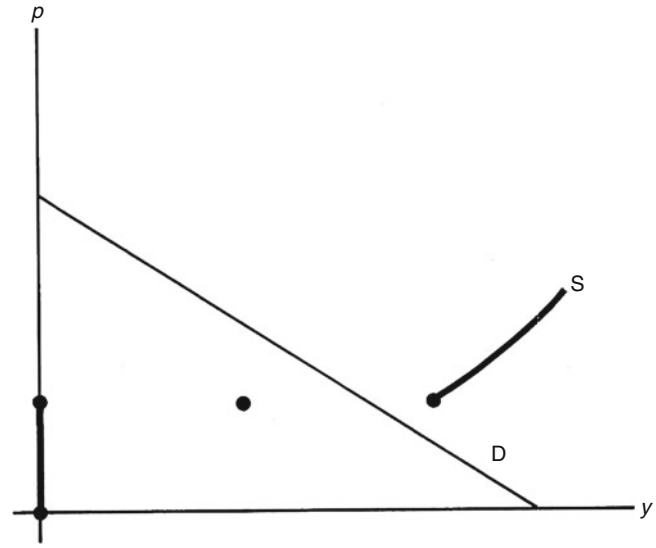
Explaining the Number of Firms

The previous discussion suggests that imperfect competition will prevail under economies of scale. On the other hand, one would expect the number of firms in an industry to be inversely related to the degree of scale economies relative to the extent of the market. Novshek's (1980) approach yields a rigorous version of this idea. Novshek considers a model of Cournot oligopoly with free entry where potential entrants adopt themselves the Cournot conjecture that active firms will keep their output constant (as proposed by Bain 1956; Sylos-Labini 1962).

An example will illustrate Novshek's method. Consider a market where the aggregate (inverse) demand is given by $p = a - b\,Y$. The number of firms in the industry is not given, but all firms, incumbent or potential, have access to the same cost function: $C(y) = g + cy$. The positive parameter g is a set-up cost: the larger g, the stronger the economies of scale. Similarly, the larger a or the smaller b, the larger the extent of the market.

Write $\hat{y}(n) = (a - c)/(n + 1)b$. This is the output of a firm at the (unique and symmetric) Cournot equilibrium for n active firms. For this

Economies and Diseconomies of Scale, Fig. 5



to be a Novshek (or Long-run Cournot) Equilibrium we require that: (a) the price be not less than average cost; this is a *no exit* condition. (b) No potential entrant has incentives to enter, i.e., $a - b(n\hat{y}(n) + y) \leq c + g/y$ for all $y > 0$: this is a *no entry* condition.

It can be checked that (a) and (b) impose lower and upper bounds on the number of firms n^* that can prevail at a Novshek Equilibrium. The no exit condition implies that $n^* + 1 \leq (a - c)/\sqrt{(bg)}$, and the no entry condition implies that $n^* + 1 \geq (1/2)(a - c)/\sqrt{(bg)}$. The expression $(a - c)/\sqrt{(bg)}$ can be viewed as an index of the 'extent of the market relative to the scale economies', since it is increasing in a and decreasing in b and g. Both bounds increase with this index, and in this sense the number of firms in an industry increases with it.

Perfect Competition as a Limit

Section, "Perfect Competition as Price Taking Behaviour" above discussed the existence difficulties that appear when production functions are not concave. These difficulties, serious for competitive equilibrium, are attenuated when considering noncompetitive equilibria: existence results for general equilibrium models can be found in Arrow and Hahn (1971) and Silvestre (1977, 1978). We focus now on partial equilibrium markets of the Novshek type (as in section, "Explaining the Number of Firms" above, but perhaps with U-shaped costs). Consider a sequence of such markets each with the same technology but with increasing size of the consumer sector (say, the parameter b of section, "Explaining the Number of Firms" above tends to zero). Equilibria turn out to exist at least for all but a finite number of markets in such a sequence.

Moreover, the equilibria of such a sequence converge to an optimal state (zero welfare loss) where the price equals the marginal cost. Such a limiting state can motivate an alternative definition of a (long-run) competitive equilibrium: this approach has the virtue of providing a justification (from the noncooperative, Cournot viewpoint) of the price taking postulate (see Mas-Colell 1980, 1981).

The convergence to long-run competitive equilibrium obtains both in the case of U-shaped cost curves and in the case of everywhere decreasing average costs (see Guesnerie and Hart 1985). This suggests a certain degree of compatibility between increasing returns and competition, in the sense that if the economy is sufficiently large, the price will be approximately equal to marginal cost in either case (even when price taking behaviour is ruled out). But Guesnerie and Hart also show that

the per capita welfare loss tends to zero much more rapidly in the U-cost case. Thus, in their words, 'there remains a sense in which everywhere increasing returns do cause greater problems for the competitive model than do increasing returns which are eventually exhausted' (p. 541).

Normative Analysis

Consider a commodity that is produced with economies of scale and that is sold in a market at a uniform price. Efficiency requires that price be equal to marginal cost. But the marginal cost is lower than the average cost. Hence, efficient pricing requires that the producing firm suffer losses. This is a basic obstacle to efficiency under increasing returns to scale. (When the commodity is not easily transferable among buyers efficiency can sometimes be achieved by means of price discrimination or nonlinear pricing schedules.)

One institutional arrangement that can in principle resolve the conflict is the public ownership of the firm. The firm can then be instructed to set prices equal to marginal costs and be subsidized for the resulting losses (see Hotelling 1938). This motivates the concept of Marginal Cost Pricing Equilibrium (see Guesnerie 1975; Beato 1982), an extension of the notion of general competitive equilibrium where firms with increasing returns to scale must follow the marginal cost pricing rule instead of profit maximization. Any efficient allocation can be attained as a marginal cost pricing equilibrium for some redistribution of income. But setting prices equal to marginal cost only guarantees the first order conditions for efficiency, not sufficient here. Thus, one should not expect that all marginal cost pricing equilibria will be efficient. A weaker desideratum is the existence of at least one efficient marginal cost pricing equilibrium compatible with a given income distribution. This turns out to obtain in some special cases (e.g., when there is a representative consumer) but not in general (see Guesnerie 1975; Brown and Heal 1979, 1980; Beato and Mas-Colell 1985).

There are, on the other hand, practical obstacles to the achievement of efficiency by a publicly owned firm. First, the absence of a profit maximization target may reduce the incentives for cost minimization. Second, the implied redistribution from taxpayers to buyers may be ethically objectionable if, say, only the wealthy are buyers. Or the public firm may find itself legally or politically constrained to break even, because it may in practice be hard to distinguish the losses mandated by marginal cost pricing from those caused by mismanagement. First best efficiency is then unattainable.

An interesting second best problem for a publicly owned or regulated firm with economies of scale is the following one (see Ramsey 1927; Boiteux 1956). Let the firm be constrained to break even, and let it sell the same commodity in two separate markets. Efficiency would require setting the prices in both markets equal to the (common) marginal cost, but this violates the break-even constraint. The second best solution requires charging different prices: the market with less elastic demand is charged a higher price.

See Also

- External Economies
- Fixed Factors
- Increasing Returns to Scale
- Indivisibilities
- Internal Economies
- Returns to Scale

Bibliography

Arrow, K.J. 1979. The division of labor in the economy, the polity and society. In *Adam Smith and modern political economy*, ed. Gerald P. O'Driscoll Jr. Ames: Iowa State University Press.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Bain, J.S. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1982. *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Beato, P. 1982. The existence of marginal cost pricing with increasing returns. *Quarterly Journal of Economics* 97: 669–689.

Beato, P., and A. Mas-Colell. 1985. On marginal cost pricing with given tax-subsidy rules. *Journal of Economic Theory* 37: 356–365.
Boiteux, M. 1956. Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica* 24: 22–40.
Brown, D.J., and G. Heal. 1979. Equity, efficiency and increasing returns. *Review of Economic Studies* 46: 571–585.
Brown, D.J., and G. Heal. 1980. Two-part tariffs, marginal cost pricing and increasing returns in a general equilibrium model. *Journal of Public Economics* 13: 25–49.
Chipman, J.S. 1965. A survey of the theory of international trade: Part 2, the neo-classical theory. *Econometrica* 33: 685–760.
Chipman, J.S. 1970. External economies of scale and competitive equilibrium. *Quarterly Journal of Economics* 84(3): 347–385.
Fuss, M., and D. McFadden (eds.). 1978. *Production economics: A dual approach to theory and applications*. Amsterdam: North-Holland.
Guesnerie, R. 1975. Pareto optimality in non-convex economies. *Econometrica* 43: 1–29.
Guesnerie, R., and O. Hart. 1985. Welfare losses due to imperfect competition: Asymptotic results for Cournot–Nash equilibria with and without free entry. *International Economic Review* 26(3): 525–545.
Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.
Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.
McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 54–71.
Mas-Colell, A. (ed.). 1980. *Noncooperative approaches to the theory of perfect competition*. Symposium issue, *Journal of Economic Theory.* Reprinted, New York: Academic Press, 1982.
Mas-Colell, A. 1981. Cournotian foundations of Walrasian equilibrium theory: An exposition of recent theory. In *Advances in economic theory*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.
Novshek, W. 1980. Cournot equilibrium with free entry. *Review of Economic Studies* 47: 473–486.
Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.
Scherer, F.M. 1980. *Industrial market structure and economic performance*, 2nd ed. Boston: Houghton Mifflin.
Silvestre, J. 1977. General monopolistic equilibrium under non-convexities. *International Economic Review* 18: 425–434.
Silvestre, J. 1978. Increasing returns in general non-competitive analysis. *Econometrica* 46(2): 397–402.
Sylos-Labini, P. 1962. *Oligopoly and technical progress*. Trans. Elizabeth Henderson, Cambridge, MA: Harvard University Press.

Economy as a Complex System

Alan Kirman

Abstract

Complex systems are composed of particles or agents which interact directly with each other. The rules for this interaction may be very simple and may not reflect the sort of rationality associated with standard economic models. Interaction is not through some exogenously given market, nor does it depend on the complicated reasoning involved in game theory. A complex system exhibits emergent aggregate properties as it organizes itself, and these can explain important phenomena such as bubbles, herding behaviour, and segregation. In each case the aggregate state of the economy or market could not be predicted from the average behaviour of the individuals.

Keywords

Agent-based models; Bubbles; Chaos; Clusters; Complex systems; Complexity; Computational complexity; Coordination; Deterministic chaos; Economy as a complex system; Econophysics; Emergence; Equilibrium; Ergodicity; Evolution; Excess volatility; Financial market contagion; Forecasting; Game theory; Herding; Imperfect competition; Law of large numbers; Minority game; Neighbourhood effects; Prisoner's Dilemma; Punctuated equilibria; Residential segregation; Social interaction; Social networks; Statistical mechanics; Statistical physics; Steady state; Tit for tat; Tournaments

JEL Classifications

D85

Introduction

The term 'complex system' has been widely used in science and many different definitions have

been given. Frequently, rather than give a definition of such a system, scientists have fallen back on certain characteristics that these systems exhibit. For example, emergence, self-organization, synergetics, collective behaviour, and non-equilibrium have all been cited in this regard. It is useful at the outset to make the distinction between 'complexity' and 'complex system'. The former involves a number of ideas which are important in economics and which are inherited from computer science but which will not be dealt with here. In particular, the notion of computational complexity, as it applies to decision-making or to the computation of equilibria or of dynamic programming problems is central to certain aspects of economic theory. However, here the discussion will turn on the idea of the economy as a complex, adaptive, evolving system. For economists the first real incarnation of this approach was with the introduction of deterministic chaos. The idea of complex dynamic behaviour, which would not explode or cycle or converge to a steady state, was fascinating for a science long dominated by the ideas of convergence to a static equilibrium or to a steady state. Jean-Michel Grandmont (1985) developed a simple model of 'business cycles' involving the 'tent map', which gave rise to such chaotic behaviour. Apart from the idea of the complicated dynamics involved, it was clear that the fact that a small perturbation in the initial conditions governing such a process could produce radically different trajectories was also of great intellectual interest. Two important innovations were involved. Firstly, there was the idea that the economy should be thought of as a truly dynamic system and that the initial conditions of such a system might play a key role. Secondly, there was the idea that there might be no continuity in the dependence on those initial conditions and that small changes might radically influence the trajectory of the system; hence the famous allusion to the influence of the fluttering of a butterfly's wing on the world's weather. These two aspects led economists to focus their attention on deterministic chaos. Yet in making such a close link between complexity and chaos, economists may have lost sight of the broader implications of complexity for the analysis of economic systems.

To see why this is so, consider what sort of systems are referred to as 'complex' in other disciplines. Typically they have some, or all, of the following characteristics:

- The agents are heterogeneous and interact directly with each other.
- The interaction and the information of agents are 'local'.
- The agents' behaviour is governed by simple 'rules of thumb'.
- The aggregate behaviour of the system is not that of an 'average' or representative agent.
- This aggregate behaviour 'emerges' from the complicated interaction between the individuals.

To someone who has not studied theoretical economics, all of these characteristics might seem rather intuitive as features of an economy. Yet they are very different from the traditional view. In that view, the economy is a system in which the only interaction is through the market. By this it is meant that agents react to signals from some central authority such as an auctioneer. In some way the central prices adjust so as to coordinate the activities of the agents. The system adjusts in this way until the activities are coordinated – for example, in a market economy, until aggregate demand for all products is equal to the aggregate supply of those products. Once this is achieved, the signals will not change and no agent has an incentive to modify his behaviour and to deviate from this 'equilibrium state'.

This description reveals another important feature of the collective model. No agent takes account of any influence that he might have on the outcome of the system. Many economists will react to this description by arguing that models of 'imperfect competition' abound, and in these models agents take into account the impact of their actions on the state of the system and know that other agents do the same.

This brings us to a second view of the economy, that based on game theory. Here all agents take account of the reciprocal impacts of their

actions and know that all the other agents do the same. This view is very different from the basic model of the economy. However, it is also very different from that of a complex system, since it attributes unlimited calculating capacity and depth of reasoning to the agents.

The vision of the economy as a complex system falls between these two approaches. It requires neither the central coordinating mechanism of the competitive market, nor the analytically sophisticated players of game theory.

A good comparison might be between, on the one hand, an economy organized as a set of markets that are open simultaneously, each with an auctioneer and, on the other, an ants' nest. In the former there are structured central price-giving mechanisms, and the actors gradually reveal their willingness or non-willingness to pay until the goods are allocated efficiently. In the latter, the individuals pursue their own different activities and react to each other and to outside stimuli. The system organizes itself but there is no central mechanism for achieving such organization. No one would think of trying to describe the activity of an ants' nest by examining the behaviour of the 'representative ant' yet many would describe the allocation of effort and resources as 'efficient'.

The sort of system that could be described as complex in the sense outlined above can be physical or biological or social. A typical reaction to the use of physical or biological analogies in economics or other social sciences is that social systems are populated by individuals that have intentions and undertake purposeful activity, while the other systems are composed of purposeless molecules or particles. It is therefore argued that the sort of analysis that can be applied to the other systems is not pertinent to the analysis of economic systems. This reasoning does not stand up to close inspection. If individuals follow well-defined rules and their interaction is well specified, the simple models that are used in physical and biological models can be applied. Precisely why the individuals should follow these rules is a different question.

Why is the complex systems approach of particular interest currently to economists? Economic theory has recently been attacked on two fronts. The first is the problem of aggregation: how is the behaviour of the economic system related to that of the individuals that make it up? The second is the question of why individuals behave as they do. The answer to the first question is simple but undermines much of modern macroeconomics that is based on the idea that the behaviour of the aggregate can be treated as the behaviour of an individual. Yet what is known is that the standard model of a system composed of isolated individuals each solving his own maximizing problem does not allow one to treat the system as an individual. (This is not the place to enter into the details of this assertion but the basic argument is given in Kirman 1992, and stems from the results of Sonnenschein, Mantel and Debreu). The second question is that posed by behavioural economics that questions the idea of the isolated maximizing individual. Ideas from Simon (1957) onwards have suggested that individuals reason in a limited and local way. Experiments, observation, and examination of the neural processes utilized in making decisions all suggest that *homo economicus* is not an accurate or adequate description of human decision making. (For a good survey of the relevant literature, see Rabin 1998).

All of this suggests that one might want to take a very different view of how the economy functions. In particular, the notion of a complex system as used in many parts of science seems to correspond well to an intuitive vision of the economy. Just as in an ants' nest individuals perform tasks without having any idea of the behaviour of the system, individuals in an economy go about their business and achieve a remarkable degree of coordination. Take a simple example that of bees in a hive. The tasks for house bees are varied but temperature control is one of the important duties. When the temperature is low, bees cluster to generate heat for themselves, but when it is high some of them fan their wings to circulate air throughout the hive. The general hive temperature required is between 33° and 36°C, while the brood chamber requires a constant heat of 35°. Honey has to be cured in order to ripen, and this also requires the help of circulating air. According to Crane (1999), 12 fanning bees positioned across a hive entrance 25 cm wide can produce an air flow amounting to

50–60 litres per minute. This fanning can go on day and night during the honey-flow season. Honeybees' wings beat 11,400 times per minute, thus making their distinctive buzz.

What is the lesson here for us? The typical economist's response to this phenomenon would be to consider a representative bee and then study how its behaviour responds to the ambient temperature. This would be a smooth function of temperature, wing beats going up or down with the temperature. Yet this is not what happens at all. Bees have different threshold temperatures and they are either on (beating at 11,400 beats per minute) or off. As the temperature rises more bees join in. Thus collectively with very simple 1, 0 rules the bees produce a smooth response. This sort of coordination, with each agent doing something simple, can only be explained by having a distribution of temperature thresholds across bees. Aggregation of individuals with specific local and differentiated behaviour produces smooth and sophisticated aggregate behaviour.

Nobody would argue that, in social systems, all coordination is achieved by simple interaction. Markets make a powerful contribution to economic coordination. Yet the important question is not whether such mechanisms exist, but how they come into being and develop and modify their rules. As already explained, the idea that the existence of such markets facilitates the allocation of resources is clear and generally accepted. What is not so clear is that the abstract idea of a market governed by centralized prices which are adjusted to equilibrate the market has any descriptive value. The idea of markets and networks of communication and transactions as emergent and changing phenomena is much more persuasive.

Considering the economy in this light is far from a new idea. When Adam Smith discusses the 'invisible hand' some of these notions are apparent, Pareto's work contains some of these ideas and Hayek is perhaps he who was closest to this vision. Schelling in his *Micromotives and Macrobehavior* (1978) clearly foresaw the role of self-organization. A recent development of these ideas had an introduction on the formal level by Foellmer (1974), who adopted the basic Ising model. He posited a system in which individuals were situated in space and whose preferences were dependent on those around them. This dependence was stochastic, that is, the probability of having certain preferences depended on the preferences of an individual's neighbours. If all the preferences are independently drawn, then one can determine the expected values of the equilibrium prices. However, if the interdependence of the individuals is too strong, this is no longer true. The 'law of large numbers' no longer applies. There is no easy transition from the micro to the macro level by simple averaging.

Foellmer's contribution was left to one side for a long time. However, the complexity approach to economics took on new life with the work at the Santa Fe Institute of a number of economists, physicists and other scientists such as Arthur, Bak, Blume, Durlauf, Geanakoplos, and Holland. A good picture of this sort of work can be found in *The Economy as an Evolving Complex System* (Anderson et al. 1988) and the two additional volumes that followed it (Arthur et al. 1997; Blume and Durlauf 2006).

The emphasis on the increasing 'socialization' of economics, which is intrinsic to models of interacting agents, permits one to introduce the influence of neighbours and groups on individual behaviour. Such an approach is standard in sociology and anthropology but has remained a very thinly populated field in economics. A good survey of this work is to be found in Durlauf and Young (2001).

One important part of the research on complex systems in economics has been that on *agent-based* models. Here the idea is to look at a set of linked individuals whose behaviour is influenced by and which influences their neighbours, and to simulate the dynamics of that interaction. Perhaps the best-known early example of this was Axelrod's work on the Prisoner's Dilemma, which is summarized in Axelrod (1997). He started from a series of tournaments. The strategies used for these were those that individuals proposed for a repeated Prisoner's Dilemma game. These strategies were then played against each other in a series of tournaments and the winning strategy turned out to be 'tit for tat', which is basically cooperative.

Axelrod was concerned that those who had entered his tournament had already anticipated the strategies that would be proposed by others. To overcome this he ran simulations in which new strategies were introduced into the pool of existing strategies. To do this he assigned existing strategies randomly to his artificial agents and then modified them using a 'genetic algorithm'. (For an introduction to the theory and use of genetic algorithms see Mitchell 1996). The set of strategies thus evolved in two ways. After the strategies had played against each other a new generation with more of the successful strategies was created. To these were also added new strategies generated by mutations and crossovers from the current population. After a while reciprocating strategies – that is, strategies which respond to cooperation with cooperation but which defect in the face of defection – took over, giving high payoffs. Here we have a selection process working on strategies that evolved rather than were consciously chosen. The behaviour of this basic but complex system – indeed, Axelrod refers to himself as a complexity scientist – led to the evolution of interesting aggregate characteristics. In this context it is also interesting to look at the work of Lindgren (1991), who also allowed the evolution of the strategy pool and generated periods of stability in which one strategy dominated, followed by periods of instability as the population was invaded by another strategy. This corresponds to the idea of 'punctuated equilibria' introduced into evolutionary theory by Eldredge and Gould (1972).

The notion of evolution, which can also be interpreted in the human or social context as adaptive learning, is important here. We can think of selection among a population of automata endowed with single strategies or of the idea that individuals learn to use more successful strategies.

Phase Transitions

Recalling the characterization of complex systems given above, it is worth considering a few examples.

In complex systems governed by local interactions, it may be the case that as a result of some perturbation there is a major change in aggregate behaviour. This is an important idea which is central to *statistical mechanics*. The idea here is that local interaction can generate a rapid transition from one 'phase' to another of an economic system and, more importantly, that one cannot simply apply the 'law of large numbers' to evaluate the impact of stochastic shocks. An example of this is provided by Bak et al. (1993), who consider a model of 'self-organized criticality' to describe an economy composed of a large number of productive units, each supplying a limited number of customers and, in turn, each supplied by a limited number of suppliers; both customers and suppliers are located near the productive unit.

The graph outlining the location of productive units is a cylindrical lattice. In other words, each production unit is supplied by the firms above it on a vertical line and supplies the customers next to it on a horizontal line. The demand for each final good producer is characterized by stochastic fluctuations, which affects the variability of orders received by the suppliers. Such orders (and shocks) are locally and vertically correlated, as every final producer is supplied by the two upstream firms situated a line up along the network representing the productive system. In such a context, characterized by local interaction, Bak et al. (1993) prove that, if individual costs are non-convex, the aggregation of small independent individual shocks may lead to large aggregate fluctuations in the productive system, breaking therefore the law of large numbers. These small shocks do not cancel each other out but are amplified by their interaction. Thus fluctuations at the aggregate level cannot be explained by reducing the whole model to one of an individual.

Coordination: The Schelling Model

Now let us pursue the discussion of the relationship between aggregate and individual behaviour. One of the important features of complex systems is that the system can coordinate on a solution which could not be predicted from a careful analysis of

the average or typical individual. In other words, patterns at the aggregate level can emerge as the individuals in an economy or market interact with each other. The emergence of such aggregate patterns cannot be forecast from the specification of the individual characteristics. A good example of this was provided by Tom Schelling at the end of the 1960s (for a summary see Schelling 1978). He introduced a model of segregation involving local interaction, in the sense that peoples' utility depends on the race of their neighbours. He showed that, even if people have only a very mild preference for living with neighbours of their own colour, as they move to satisfy their preferences complete segregation will occur.

The basic model is very simple. Take a large chess board, and place a certain number of black and white counters on the board, leaving some free places. A counter prefers to be on a square where half or more of the counters in his Moore neighbourhood, (the eight squares around him) are of its own colour (utility 1) to the opposite situation (utility 0). From the counters with utility zero, one is chosen at random and moves to a preferred location. This model, when simulated, yields complete segregation even though people's preferences for being with their own colour are not strong. Indeed, the result holds when individuals are happy even when more than half of their neighbours are of a colour different from their own. This result was greeted with surprise and has generated a large literature.

In fact, this result is not surprising and some simple physical theory (see Vinkovic and Kirman 2006), can explain the segregation phenomenon. Numerous variants on Schelling's original model have been developed. In particular, the form of the utility function used by Schelling, the size of neighbourhoods, the rules for moving, and the amount of unoccupied space have all been studied (see Pancs and Vriend 2007, for a survey). The physical model encompasses all of these variants.

An attempt to provide a formal structure has been made by Pollicott and Weiss (2001). They however, examine the limit of a Laplacian process in which individuals' preferences are strictly increasing in the number of like neighbours. In this situation it is intuitively clear that there is a strong tendency to segregation. Yet Schelling's result has become famous because the preferences of individuals for segregation were not particularly strong. The model is of interest because it illustrates the emergence of an aggregate phenomenon which is not directly foreseen from individual behaviour and because it concerns an important economic problem, that of segregation.

The physical analogue to Schelling's model, developed in Vinkovic and Kirman (2006), exhibits three features of the resultant segregation. The first is the organization of the system into 'regions' or clusters, each containing individuals of only one colour. Second, it explains the shape of the frontier between the regions. Lastly, in the case where several clusters of one colour may form it allows one to analyse the size distribution of the clusters.

The basic idea is simple. Think of utility as the negative of energy. Particles with high energy in the physical system correspond to individuals with low utility in the social system. Where are the unhappy or high-energy individuals to be found? Clearly they are individuals on the frontiers of clusters. Those within clusters of their own colour are happy and have no possibility of increasing their utility by moving. Those on the frontier, on the other hand, are in contact with those of the other colour and there may be too many of the latter. In this case these individuals correspond to particles with high energy. A physical system with these characteristics will seek to minimize its energy. The energy is highest on the frontier between clusters. Thus the way for the system to minimize its energy is to reduce the length of these frontiers. It will achieve this by organizing itself into clusters, and the shape and size of these clusters will depend on the precise variant of the model. In the original model the system will organize itself into two giant clusters, each composed of individuals of one colour. If we only allow people to move to currently free places, then the number of these will be important for the outcome. If there are not enough, the system will 'freeze' with many small clusters. If, on the other hand, individuals can swap places the system will segregate, but there will be perpetual movement within it. Thus, a simple physical

model generates the result obtained by Schelling and, furthermore, shows how the form of the segregation depends on the exact version of the model. (For a discussion of the emergent properties of the Schelling model see emergence).

The 'El Farol Bar'

Another interesting example of emergent coordination is that provided by Brian Arthur (1994) in his 'El Farol Bar' problem. The simple model that he develops and which has been taken up by many physicists under the name of 'the minority game' shows how individuals using rules -of thumb can come to coordinate in a way which yields a satisfactory social outcome even though no individual had any such intention. The idea is that the bar can hold 100 people. Being at the bar with fewer than 60 people is, by common consent, better than staying at home. However once attendance goes over 60 the bar becomes too crowded and home is the preferable alternative. The question then is how people will decide whether to go to the bar. Suppose that they all reason strategically. In this case they must decide in function of what their neighbours will decide. Thus, to anticipate whether there will be more than 60 people at the bar they must reflect on the strategies employed by the others. However, they must also take into account that the others are doing the same and know that the others know that they know that they are behaving in this way. This leads to an infinite regress that poses logical problems for the foundations of such game-theoretic reasoning. Rather than attribute such calculating capacities to his agents, Brian Arthur imagined that each was endowed with a set of forecasting rules based on previous attendance at the bar. Given his set of rules the individual chooses that rule which has forecast best up to the present, 'best' meaning the forecast that has the smallest sum of squared prediction errors, for example. Now, each agent uses, as information, just the attendance observed at the bar, and updates in consequence. There is no coordinating mechanism, yet the model quickly settles to the 'equilibrium' solution with 60 people at the bar with occasional small deviations. Furthermore, each agent receives a fixed number of forecasting rules, some of which may be rather stupid. Nevertheless, coordination is achieved at the aggregate level.

Some things about this model are worth noting. It is not guaranteed that all agents will learn to forecast correctly; some may persist in erroneous forecasts. The way in which the model is set up means that whenever attendance goes to 61 many people are unhappy, which is not the case when it goes to 59. This asymmetry does not prevent the achievement of collective coordination, however. Thus, the relation between satisfactory performance at the aggregate level and satisfaction at the individual level is tenuous. While many may find this example intriguing, one might enquire as to how it can be directly applied to economic problems. An interesting answer is to be found in a book by some Oxford physicists who specialize in complex systems and who apply the model to financial markets (see Johnson et al. 2003, pp. 81–136).

Financial Markets

This brings us to another important example, that of financial markets. Models of economies with interacting agents in the spirit of complex systems may, as we have just seen, be able to show how certain aggregate coordination may emerge. They may also help us to analyse some of the observed features of markets which normal economic analysis has difficulty explaining. For example, one of the major problems with the standard model of financial markets is that they do not reproduce certain well-established stylized facts about empirical price series. In standard models, where there is uncertainty about the evolution of prices, the usual way of achieving consistency is to assume that agents have common and 'rational' expectations. Yet, if agents have such common expectations, how can there be trade? Indeed there are many 'no trade' theorems for such markets. How, then, do we deal with the fact that the volume of trade on financial markets is very important and that agents do, in fact, differ in their opinions and forecasts and that this is one

of the main sources of such trade? There is also an old problem of 'excess volatility,' that is, prices have a higher variance than the returns on the assets on which they are based. One answer is to allow for direct interaction between agents other than through the market mechanism. Models reminiscent of the Ising model from physics have been used to doing this. For example, one might suggest that individuals may change their opinions or forecasts as a function of those of other agents. In simple models of financial markets such changes may be self-reinforcing. If agents forecast an increase in the price of an asset and others are persuaded by their view, the resultant demand will drive the price up, thereby confirming the prediction. However, the market will not necessarily 'lock on' to one view for ever. Indeed, under certain rather reasonable assumptions, if agents make stochastic rather than deterministic choices, then it is certain that the system will swing back to a situation in which another opinion dominates. The stochastic choices are not irrational, however. The better the results obtained when following one opinion, the higher is the probability of continuing to hold that opinion.

Such models will generate swings in opinions, regime changes and 'long memory', all of which are hard to explain with standard analysis. An essential feature of these models is that agents are wrong for some of the time, but whenever they are in the majority they are essentially right. Thus they are not systematically irrational. (For examples of this sort of model see, Lux and Marchesi 1999; Brock and Hommes 1997; and Kirman and Teyssiere 2005, and for a recent survey, De Grauwe and Grimaldi 2006). Thus the behaviour of the agents in the market cannot correctly be described as 'irrational exuberance', in the well-known words of Alan Greenspan, Chairman of the Board of Governors of the Federal Reserve from 1987 to 2006.

Economists faced with this sort of model are often troubled by the lack of any equilibrium notion. The process is always moving; agents are neither fully rational nor systematically mistaken. Worse, the process never settles down to a particular price even without exogenous shocks. Suppose that we accept this kind of model: can we say anything analytic about the time series that result? If we consider some of these models, for certain configurations of parameters they could become explosive. There are two possible reactions to this. Since we will never observe more than a finite sample, it could well be that the underlying stochastic process is actually explosive, but this will not prevent us from trying to infer something about the data that we observe. Suppose, however, that we are interested in being able, from a theoretical point of view, to characterize the long-run behaviour of the system. In particular, if we treat the process as being stochastic and do not make a deterministic approximation, then we have to decide what, if anything, constitutes an appropriate long-run equilibrium notion. Such a concept provides an answer to those who consider that complex systems, by their nature, are not amenable to formal analysis. Foellmer et al. (2005), examined the sort of price process discussed here and produced some analytical results characterizing the process. Furthermore, they provided a long-run equilibrium notion that is not the convergence to a particular price vector.

If prices change all the time, as they will do in an evolving complex system, how may one speak of 'equilibrium'? The idea is to look at the evolving distribution of prices and to try to characterize its long-run behaviour. Foellmer et al. (2005) examined the process governing the evolution of asset prices and the profits made by traders, and gave conditions under which it is ergodic, that is, the proportion of time that the price takes on each possible value converges over time and that the *limit distribution* is unique. (For a discussion of the mathematical background, see ergodicity and nonergodicity in economics). This means that, unlike the 'anything can happen' often associated with deterministic chaos, in the long run the price and profits process does have a well-defined structure.

Conclusion

To view the economy as a complex system implies a fundamental rethinking of theoretical

economics. The basic idea is that of a decentralized system with no central source of signals, whose aggregate behaviour cannot be reduced to that of an individual. Furthermore, the individuals are endowed with local information and interact directly with each other, and their behaviour can be characterized by simple rules. Such a vision is far from new in economics. Its origins can be traced back at least to Adam Smith and a long chain of economists leads from him to Hayek and Simon, who preceded the developments described here. The most recent contributions borrow heavily from other disciplines such as statistical physics and the appearance of 'econophysics' represents a shift from the path that led from classical mechanics to axiomatic mathematical models as the basic paradigm of economic theory. This sort of approach has already allowed economists to analyse problems such as contagion, neighbourhood effects, financial bubbles, and herding behaviour, none of which fits well into the standard economic framework. In addition, many of the features that are imposed on standard models emerge as a result of the interaction between agents (see emergence).

Perhaps, most importantly, looking at economies in this way provides a very different and more intuitive vision of the economy as a vast interactive system whose aggregate properties reflect the self-organization of the system and its continual adaptation. However, entrenched ideas die hard and it remains to be seen whether Steven Hawking's prediction that the 21st century will be the 'age of complexity' will hold true for economics.

See Also

- ▶ Emergence
- ▶ Ergodicity and Nonergodicity in Economics
- ▶ Interacting Agents in Finance
- ▶ Social Interactions (Empirics)
- ▶ Social Interactions (Theory)
- ▶ Social Networks, Economic Relevance of
- ▶ Statistical Mechanics

Bibliography

Anderson, P.W., K.J. Arrow, and D. Pines. 1988. *The economy as an evolving complex system*. Redwood City: Addison-Wesley.

Arthur, W.B. 1994. Inductive reasoning and bounded rationality. *American Economic Review* 84: 406–411.

Arthur, W.B., S.N. Durlauf, and D.A. Lane, ed. 1997. *The economy as an evolving complex system II*. Reading: Addison-Wesley.

Axelrod, R. 1997. *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton: Princeton University Press.

Bak, P., K. Chen, J. Scheinkman, and M. Woodford. 1993. Aggregate fluctuations from independent sectoral shocks: Self-organized criticality in a model of production and inventory dynamics. *Ricerche Economiche* 47: 3–30.

Blume, L., and S.N. Durlauf. 2006. *The economy as an evolving complex system III*. Oxford: Oxford University Press.

Brock, W.A., and C. Hommes. 1997. A rational route to randomness. *Econometrica* 65: 1059–1095.

Crane, E. 1999. *The world history of beekeeping and honey hunting*. London: Duckworth.

De Grauwe, P., and M. Grimaldi. 2006. *The exchange rate in a behavioral finance framework*. Princeton: Princeton University Press.

Durlauf, S.N., and H.P. Young. 2001. *Social dynamics: Economic learning and social evolution*. London: MIT Press.

Eldredge, N., and S.J. Gould. 1972. Punctuated equilibria: An alternative to phyletic gradualism. In *Models in paleobiology*, ed. T.J.M. Schopf. San Francisco: Freeman, Cooper and Co.

Foellmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1: 51–62.

Foellmer, H., U. Horst, and A. Kirman. 2005. Equilibrium in financial markets with heterogeneous agents: A new perspective. *Journal of Mathematical Economics* 41: 123–155.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1046.

Johnson, N., P. Jeffries, and P.M. Hui. 2003. *Financial market complexity*. Oxford: Oxford University Press.

Kirman, A. 1992. What or whom does the representative individual represent? *Journal of Economic Perspectives* 6(2): 117–136.

Kirman, A., and G. Teyssiere. 2005. Testing for bubbles and change points. *Journal of Economic Dynamics and Control* 29: 765–799.

Lindgren, K. 1991. Evolutionary phenomena in simple dynamics. In *Artificial life II*, ed. C.G. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen. Redwood City: Addison Wesley.

Lux, T., and M. Marchesi. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397: 498–450.

Mitchell, M. 1996. *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.
Pancs, R., and N.J. Vriend. 2007. Schelling's spatial proximity model of segregation revisited. *Journal of Public Economics* 91: 1–24.
Pollicott, M., and H. Weiss. 2001. The dynamics of Schelling-type segregation models and a nonlinear graph Laplacian variational problem. *Advances in Applied Mathematics* 27: 17–40.
Rabin, M. 1998. Psychology and economics. *Journal of Economic Literature* 36: 11–46.
Schelling, T. 1978. *Micromotives and macrobehavior*. New York: W.W. Norton and Co.
Simon, H.A. 1957. *Models of man: Social and rational*. New York: Wiley.
Vinkovic, D., and A. Kirman. 2006. A physical analogue of the Schelling model. *Proceedings of the National Academy of Sciences* 103: 19261–19265.

Econophysics

J. Barkley Rosser Jr.

Abstract

Econophysics, a term neologized only in 1995, refers to physicists studying economics problems using conceptual approaches from physics. Certain ideas are emphasized, especially the ubiquity of scaling laws in distributions of financial returns, income and wealth, firm sizes, city sizes, and other economic phenomena. However, economists have been using many of these techniques since much earlier, and the influence of ideas from physics on economics dates as far back as 1801 at least. Arguably, if economics successfully absorbs the most useful of this work, 'econophysics' may cease to exist.

Keywords

Bachelier, L.; Black–Scholes formula; Bounded rationality; Brownian motion; Canard, N.-F.; Econobiology; Econochemistry; Econophysics; Lévy distribution; Lognormal distribution; Pareto distribution; Pareto, V.; Random walk; Scaling laws; Statistical mechanics

JEL Classifications

C6; D5

According to Bikas Chakrabarti (2005, p. 225), the term 'econophysics' was neologized in 1995 at the second Statphys-Kolkata conference in Kolkata (formerly Calcutta), India, by the physicist H. Eugene Stanley, who was also the first to use it in print (Stanley 1996). Mantegna and Stanley (2000, pp. viii–ix) define 'the multidisciplinary field of econophysics' as 'a neologism that denotes the activities of physicists who are working on economics problems to test a variety of new conceptual approaches deriving from the physical sciences'.

The list of such problems has included distributions of returns in financial markets (Mantegna 1991; Levy and Solomon 1997; Bouchaud and Cont 1998; Gopakrishnan et al. 1999; Sornette and Johansen 2001; Farmer and Joshi 2002), the distribution of income and wealth (Drăgulescu and Yakovenko 2001; Bouchaud and Mézard 2000; Chatterjee et al. 2005), the distribution of economic shocks and growth rate variations (Bak et al. 1993; Canning et al. 1998), the distribution of firm sizes and growth rates (Stanley et al. 1996a, b; Takayasu and Okuyama 1998; Botazzi and Secchi 2003), the distribution of city sizes (Rosser 1994; Gabaix 1999), and the distribution of scientific discoveries (Plerou et al. 1999; Sornette and Zajdenweber 1999), among other problems, all of which are seen at times not to follow normal or Gaussian patterns that can be described fully by mean and variance. The main sources of conceptual approaches from physics used by the econophysicists have been from models of statistical mechanics (Spitzer 1971), geophysical models of earthquakes (Sornette 2003), and 'sandpile' models of avalanches, the latter involving self-organized criticality (Bak 1996). An early physicist to assert the essential identity of statistical methods used in physics and the social sciences was Majorana (1942).

A common theme among those who identify themselves as econophysicists is that standard economic theory has been inadequate or insufficient to explain the non-Gaussian distributions empirically observed for various of these

phenomena, such as 'excessive' skewness and leptokurtotic 'fat tails' (McCauley, 2004). With their sense of creating and developing a new science based on physics that is superior to the older conventional economics, many of the econophysicists have focused their publishing efforts in physics journals, notably *Physica A*, *Physical Review E*, and *European Physical Journal B*, to name some of the most frequently used ones, along with the general science journal *Nature* and some more clearly multidisciplinary journals such as *Quantitative Finance*. However, increasingly some of the econophysicists have begun to publish jointly with economists, with some of these papers appearing in economics journals as well. This should not be surprising in that the emergence of econophysics followed fairly shortly after the influential interactions and discussions that occurred between groups of physicists and economists at the Santa Fe Institute (Anderson et al. 1988; Arthur et al. 1997), with some of the physicists involved in these discussions also becoming involved in the econophysics movement.

Now we come to a great curiosity and irony in this matter: some of the main techniques used by econophysicists were initially developed by economists (with many others developed by mathematicians), and some of the ideas associated with economists were developed by physicists. Thus, in a sense, these efforts by physicists resemble carrying coals to Newcastle, except that it must be admitted that many economists either forgot or never knew of these issues or methods. This is true of the most canonical of such models, the Pareto distribution.

The Empirical Focus on Scaling Laws (Power Laws)

If there is a single issue that unites the econophysicists it is the insistence that many economic phenomena occur according to distributions that obey scaling laws rather than Gaussian normality. Whether symmetric or skewed, the tails are fatter or longer than they would be if Gaussian, and they appear to be linear in figures with the logarithm of a variable plotted against its cumulative probability distribution. They search for physics processes, most frequently from statistical mechanics, that can generate these non-Gaussian distributions that obey scaling laws.

The canonical (and original) version of such a distribution was discovered by the mathematical economist and sociologist, Vilfredo Pareto, in 1897. Let N be the number of observations of a variable that exceed a value x with A and α positive constants. Then

$$N = Ax^{-\alpha}. \tag{1}$$

This exhibits the scaling property in that

$$\ln(N) = \ln A - \alpha \ln(x). \tag{2}$$

This can be generalized to a more clearly stochastic form by replacing N with the probability that an observation will exceed x. Pareto formulated this to explain the distribution of income and wealth, and believed that there was a universally true value for α that equalled about 1.5. More recent studies (Clementi and Gallegati 2005) suggest that it is only the upper end of income and wealth distributions that follow such a scaling property, with the lower ends following the lognormal form of the Gaussian distribution that is associated with the random walk, originally argued for the whole of the income distribution by Gibrat (1931).

The random walk and its associated lognormal distribution is the great rival to the Pareto distribution and its relatives in explaining stochastic economic phenomena. It was only a few years after Pareto did his work that the random walk was discovered in a Ph.D. thesis about speculative markets by the mathematician Louis Bachelier (1900), five years prior to Einstein using it to model Brownian motion, its first use in physics (Einstein 1905). Although the Paretian distribution would have its advocates for explaining stochastic price dynamics (Mandelbrot 1963), the random walk would become the standard model for explaining asset price dynamics for many decades, although it would be asset returns that would be so modelled rather than asset prices

themselves directly as Bachelier did originally. As a further irony, it was a physicist, M. F. M. Osborne (1959), who was among the influential advocates of using the random walk to model asset returns. It was the Gaussian random walk that would be assumed to underlie asset price dynamics when such basic financial economics concepts as the Black–Scholes formula would be developed (Black and Scholes 1973). If we let p be price, R be the return due to a price increase, B be debt, and σ be the standard deviation of the Gaussian distribution, then Osborne characterized the dynamic price process by

$$dp = Rpdt + \sigma pdB. \quad (3)$$

Meanwhile, a variety of efforts were made over a long time by physicists, mathematicians and economists to model a variety of phenomena using either the Pareto distribution or one its relatives or generalizations, such as the stable Lévy (1925) distribution, prior to the clear emergence of econophysics. Alfred Lotka (1926) saw scientific discoveries as following this pattern. George Zipf (1941) would see city sizes as doing so. Benoit Mandelbrot (1963) saw cotton prices doing so and was inspired to discover fractal geometry from studying the mathematics of the scaling property (Mandelbrot 1983, 1997). Ijiri and Simon (1977) saw firm sizes also following this pattern, a result more recently confirmed by Axtell (2001).

Economists Doing Econophysics?

Also, economists would move to use statistical mechanics models to study a broader variety of economic dynamics prior to the emergence of econophysics as such. Those doing so included Hans Föllmer (1974), Lawrence Blume (1993), Steven Durlauf (1993), William Brock (1993), Duncan Foley (1994) and Michael Stutzer (1994), with Durlauf (1997) providing an overview of an even broader set of applications. However, by 1993 the econophysicists were fully active even if they had not yet identified themselves by this term.

While little of this work explicitly focuses on generating outcomes consistent with scaling laws, it is certainly reasonable to expect that many of them could. It is true that the more traditional view of efficient markets with all agents possessing full information rational expectations about a single stable equilibrium is not maintained in these models, and therefore the econophysics critique carries some weight. However, many of these models do make assumptions of at least forms of bounded rationality and learning, with the possibility that some agents may even conform to the more traditional assumptions. Stutzer's (1994) reconciles the maximum entropy formulation of Gibbsian statistical mechanics with a relatively conventional financial economics formulation of the Black–Scholes options formula, based on Arrow–Debreu contingent claims (Arrow 1974). Brock and Durlauf (2001) formalize heterogeneous agents socially interacting within a utility maximizing, discrete choice framework. Neither of these specifically generates scaling law outcomes, but there is nothing preventing them from doing so potentially.

While some econophysicists seek to integrate their findings with economic theory, as noted above many seek to replace conventional economic theory, seeing it as useless and limited. An irony in this effort is that it has been argued that conventional neoclassical economic theory itself was substantially a result of importing 19th-century physics conceptions into economics, with not all observers approving of this (Mirowski 1989). The culmination of this effort is seen by many as being Paul Samuelson's *Foundations of Economic Analysis* (1947), whose undergraduate degree was in physics at the University of Chicago. Samuelson himself noted approvingly that Irving Fisher's 1892 dissertation (1926) was partly supervised by the pioneer of statistical mechanics, J. Willard Gibbs (1902), and as far back as 1801 Nicholas-François Canard conceived of supply and demand ontologically being contradicting 'forces' in a physics sense. So the interplay between economics and physics has been going on for far longer and is considerably more complicated than is usually conceived.

Related Trans-disciplinary Movements

Curiously but unsurprisingly given the tremendous attention given to the new econophysics movement, it has spawned imitators since 2000 in the form of *econochemistry* and *econobiology*, although these have not had nearly the same degree of development. The former term is the title of a course of study established at the University of Ulm by Barbara Mez-Starke, and was used to describe the work of Hartmann and Rössler (1998) at a conference in 2002 in Urbino, Italy (see also Padgett et al. 2003, for a more recent effort). The latter term first appeared in Hens (2002), although McCauley (2004, pp. 196–9) dismisses it as not a worthy competitor for econophysics. Nevertheless, there has long been a tradition among economists of advocating drawing more from biology for inspiration than from physics (Hodgson 1993), going back at least as far as Alfred Marshall's famous declaration that economics is 'a branch of biology broadly interpreted' (Marshall 1920, p. 637), even as Marshall's actual analytical apparatus arguably drew more from physics than from biology.

In any case, one trend we can expect for some time is an increase in coauthoring between economists and physicists within the area of econophysics (Lux and Marchesi 1999; Li and Rosser 2004). Very likely we shall eventually see the more useful ideas of econophysics coming to be absorbed into economics proper. As that comes to pass, it may also come to pass that the separate and distinct movement we now know as econophysics will cease to exist and will be forgotten, just as most economists do not think about the physics roots of standard neoclassical economic theory today.

See Also

- ▶ Economy as a Complex System
- ▶ Evolutionary Economics
- ▶ Gibrat, Robert Pierre Louis (1904–1980)
- ▶ Gibrat's Law
- ▶ Inequality (Global)
- ▶ Inequality (Measurement)
- ▶ Lognormal Distribution
- ▶ Non-linear Time Series Analysis
- ▶ Pareto Distribution
- ▶ Pareto, Vilfredo (1848–1923)
- ▶ Power Laws
- ▶ Redistribution of Income and Wealth
- ▶ Science, Economics of
- ▶ Systems of Cities
- ▶ Transfer of Technology
- ▶ Urban Growth

Bibliography

Anderson, P., K. Arrow, and D. Pines, eds. 1988. *The economy as an evolving complex system*. Redwood City, CA: Addison-Wesley.

Arrow, K. 1974. *Essays in the theory of risk bearing*. Amsterdam: North-Holland.

Arthur, W., S. Durlauf, and D. Lane, eds. 1997. *The economy as an evolving complex system II*. Redwood City, CA: Addison-Wesley.

Axtell, R. 2001. Zipf distribution of firm sizes. *Science* 293: 1818–1820.

Bachelier, L. 1900. Théorie de la spéculation. *Annales Scientifique de l'École Normale Supérieure* III-17, 21–86. (English translation). In *The random character of stock market prices*, ed. P. Cootner, 1964. Cambridge, MA: MIT Press.

Bak, Per. 1996. *How nature eorks: The science of self-organized criticality*. New York: Copernicus Press for Springer-Verlag.

Bak, P., K. Chen, J. Scheinkman, and M. Woodford. 1993. Aggregate fluctuations from independent sectoral shocks: self-organized criticality in a model of production and inventory dynamics. *Ricerche Economiche* 47: 3–30.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5: 387–424.

Botazzi, G., and A. Secchi. 2003. A stochastic model of firm growth. *Physica A* 324: 213–219.

Bouchaud, J.-P., and R. Cont. 1998. A Langevin approach to stock market fluctuations and crashes. *European Physical Journal B* 6: 543–550.

Bouchaud, J.-P., and M. Mézard. 2000. Wealth condensation in a simple model of economy. *Physica A* 282: 536–545.

Brock, W. 1993. Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance. *Estudios Económicos* 8: 3–55.

Brock, W., and S. Durlauf. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68: 235–260.

Canard, N.-F. 1801. *Principes d'Économie Politique*, 1969. Rome: Edizioni Bizzarri.
Canning, D., L. Amaral, Y. Lee, M. Meyer, and H. Stanley. 1998. A power law for scaling the volatility of GDP growth rates with country size. *Economics Letters* 60: 335–341.
Chakrabarti, B. 2005. Econphys-Kolkata: a short story. In *Econophysics of wealth distributions*, ed. A. Chatterjee, S. Yarlagadda, and B. Charkrabarti. Milan: Springer.
Chatterjee, A., S. Yarlagadda, and B. Charkrabarti, eds. 2005. *Econophysics of wealth distributions*. Milan: Springer.
Clementi, F., and M. Gallegati. 2005. Power law tails in the Italian personal income distribution. *Physica A* 350: 427–438.
Drăgulescu, A., and V. Yakovenko. 2001. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A* 299: 213–221.
Durlauf, S. 1993. Nonergodic economic growth. *Review of Economic Studies* 60: 349–366.
Durlauf, S. 1997. Statistical mechanics approaches to socioeconomic behavior. In *The Economy as a complex evolving system II*, ed. W. Arthur, S. Durlauf, and D. Lane. Redwood City, CA: Addison-Wesley.
Einstein, A. 1905. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von der ruhenden Flüssigkeiten suspendierten Teichen. *Annalen der Physik* 17: 549–560.
Farmer, J., and S. Joshi. 2002. The price dynamics of common trading strategies. *Journal of Economic Behavior and Organization* 49: 149–171.
Fisher, I. 1926. *Mathematical investigations into the theory of value and prices*. New Haven: Yale University Press.
Foley, D. 1994. A statistical equilibrium theory of markets. *Journal of Economic Theory* 62: 321–345.
Föllmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1: 51–62.
Gabaix, X. 1999. Zipf's law for cities: an explanation. *Quarterly Journal of Economics* 114: 739–767.
Gibbs, J. 1902. *Elementary principles in statistical mechanics*. New Haven: Yale University Press.
Gibrat, R. 1931. *Les Inégalités Économiques*. Paris: Sirey.
Gopakrishnan, P., V. Plerou, L. Amaral, M. Meyer, and H. Stanley. 1999. Scaling of the distributions of fluctuations of financial market indices. *Physical Review E* 60: 5305–5316.
Hartmann, G., and O. Rössler. 1998. Coupled flare attractors – a discrete prototype for economic modelling. *Discrete Dynamics in Nature and Society* 2: 153–159.
Hens, T. 2002. Evolutionary portfolio theory. *Asset allocation almanac: special report #4*. Merrill Lynch. Online. Available at http://www.evolutionary finance.ch/uploads/media/MerrillLynch.pdf. Accessed 22 May 2006.
Hodgson, G. 1993. *Economics and evolution: bringing life back into economics*. Ann Arbor: University of Michigan Press.
Ijiri, Y., and H. Simon. 1977. *Skew distributions and the sizes of business firms*. Amsterdam: North-Holland.
Levy, M., and S. Solomon. 1997. New evidence for the power-law distribution of wealth. *Physica A* 242: 90–94.
Lévy, P. 1925. *Calcul des Probabilités*. Paris: Gauthier-Villars.
Li, H., and J. Rosser Jr. 2004. Market dynamics and stock price volatility. *European Physical Journal B* 39: 409–413.
Lotka, A. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 12: 317–323.
Lux, T., and M. Marchesi. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397: 498–500.
Majorana, E. 1942. Il valore delle leggi statistiche nelle fisica e nelle scienze sociali. *Scientia* 36: 58–66.
Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36: 394–419.
Mandelbrot, B. 1983. *The fractal geometry of nature*. San Francisco: W.H. Freeman.
Mandelbrot, B. 1997. *Fractals and scaling in finance*. New York: Springer-Verlag.
Mantegna, R. 1991. Lévy walks and enhanced diffusion in Milan stock exchange. *Physica A* 179: 232–242.
Mantegna, R., and H. Stanley. 2000. *An introduction to econophysics: correlations and complexity in finance*. Cambridge: Cambridge University Press.
Marshall, A. 1920. *Principles of economics*. 8 ed. London: Macmillan.
McCauley, J. 2004. *Dynamics of markets: econophysics and finance*. Cambridge: Cambridge University Press.
Mirowski, P. 1989. *More heat than light: economics as social physics, physics as nature's economics*. Cambridge: Cambridge University Press.
Osborne, M. 1959. Brownian motion in stock markets. *Operations Research* 7: 145–173.
Padgett, J., D. Lee, and N. Collier. 2003. Economic production as chemistry. *Industrial and Corporate Change* 12: 843–877.
Pareto, V. 1897. *Cours d'Économie Politique*. Paris and Lausanne. Trans. In *Manual of Political Economy*, ed. A. Schwier, 1971. New York: Kelly.
Plerou, V., L. Amaral, P. Gopakrishnan, M. Meyer, and H. Stanley. 1999. Similarities between the growth dynamics of university research and competitive economic activities. *Nature* 400: 433–437.
Rosser, J. Jr. 1994. Dynamics of emergent urban hierarchy. *Chaos, Solitons & Fractals* 4: 553–562.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Sornette, D. 2003. *Why stock markets crash: critical events in complex financial systems*. Princeton: Princeton University Press.
Sornette, D., and A. Johansen. 2001. Significance of log-periodic precursors to financial crashes. *Quantitative Finance* 1: 452–471.

Sornette, D., and D. Zajdenweber. 1999. Economic returns of research: The Pareto law and its implications. *European Physical Journal B* 8: 653–664.
Spitzer, F. 1971. *Random fields and interacting particle systems*. Providence: American Mathematical Society.
Stanley, H., V. Afanasyev, L. Aamaral, S. Buldyrev, A. Goldberger, S. Havlin, H. Leschhorn, P. Maass, R. Mantegna, C.-K. Peng, P. Prince, M. Salinger, M. Stanley, and G. Viswanathan. 1996a. Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Physica A* 224: 302–321.
Stanley, M., L. Amaral, S. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. Salinger, and H. Stanley. 1996b. Scaling behavior in the growth of companies. *Nature* 379: 804–806.
Stutzer, M. 1994. The statistical mechanics of asset prices. In *Differential equations, dynamical systems, and control science: A festschrift in honor of lawrence markus*, ed. K. Elworthy, W. Everitt, and E. Lee, vol. 152. New York: Marcel Dekker.
Takayasu, H., and K. Okuyama. 1998. Country dependence on company size distributions and a numerical model based on competition and cooperation. *Fractals* 6: 67–79.
Zipf, G. 1941. *National unity and disunity*. Bloomington, IN: Principia Press.

Eden, Frederick Morton (1766–1809)

K. Tribe

Keywords
Eden, F. M.; Poverty

JEL Classifications
B31

The son of Sir Robert Eden, F.M. Eden was educated at Oxford, gaining a Master's degree in 1789. A co-founder of the Globe Insurance Company, he published in 1797 the three volumes of his investigation into the conditions of the labouring poor, *The State of the Poor*. This work was perhaps the most detailed appraisal of social legislation and its actual workings that had appeared, and the findings provided ample material for ensuing debate on the best form of dealing with poverty and pauperism. In the years that followed Eden wrote a number of pamphlets on related issues.

The greater part of *The State of the Poor* records Eden's findings relating to the actual conditions prevailing in the parishes of England. Stimulated by the high prices prevailing in 1794–5, Eden initially set out to study the condition of the poor, but later extended this to the labouring classes. He encountered at times great resistance from local parish authorities, but despite this he was able to gather a considerable amount of information on wage levels, diet and prices. This was linked to an appraisal of the nutritional value of available foodstuffs, such that it was possible to arrive at some kind of comparative assessment of levels of poverty and want. It emerged from his empirical findings that the actual conditions and treatment of the poor varied greatly from parish to parish, this in part reflecting the patchwork of legislation that had grown up over the years in relation to the pauper and the workless. He argued however that existing legislation implied a policy of support for the indigent, and that in general a civilized society had an obligation to make such provision.

Selected Works

1797. F. M. Eden, *The state of the poor: Or an history of the labouring classes in England*, 3 vols. London.

Edgeworth as a Statistician

Stephen M. Stigler

Francis Edgeworth was the leading theorist of mathematical statistics of the latter half of the 19th century, though his influence was diminished by the difficulty of his exposition. He is most frequently remembered today for his work on the

Edgeworth Series, but in fact he touched on nearly every sphere of modern statistics, from the analysis of variance to stochastic models, to multivariate analysis, to the asymptotic theory of maximum likelihood estimates, to inventory theory. In some areas such as correlation, his work was decisive in the development of all that followed.

Edgeworth's first purely statistical work was published in 1883, when he began a series of papers examining the methods, rationale and philosophical foundations of probability and its application to the analysis of observational data. Most of this work appeared in the *Philosophical Magazine, Mind*, or the *Journal* of the London (later the Royal) Statistical Society. Between 1883 and 1890 he published over 30 separate papers on a wide selection of statistical topics; these works are best viewed as the tracks left by a first-rate mind as it took an excursion through territory that had already been explored. He found much that was new, but his principal occupation re-examining past works, particularly those of Laplace, to see how they might be used in social science. A major (and under appreciated) accomplishment of this period was Edgeworth's explanation of how simple significance tests could be used to compare averages. The mathematical technique was not new, but the conceptual framework was subtly different from that of the early astronomers, and while Edgeworth's (1885b) explanation may today seem elementary, it had a lasting widespread impact. In subsequent work (Edgeworth 1885c; Stigler 1978) he developed what might now be viewed as an analysis for an additive effects model for a two-way classification, and he was sensitive to the effect non-normality or serial dependence could have upon the procedures.

Edgeworth's main orientation in his inferential work was Bayesian, and he presented both philosophical and mathematical investigations of this approach. To Edgeworth, a prior distribution was based in a rough way upon experience. A uniform prior was often justified because, Edgeworth observed, we do not find a pattern in nature that tends to favour one set of values for its constants over another set. Edgeworth tempered this with a realization that inferences would frequently not be very sensitive to the prior specification (Edgeworth 1885a). When evaluating the significance of differences, however, Edgeworth reverted to a sampling theory viewpoint. One of his 1883 works includes a derivation of Student's *t*-distribution as the posterior distribution for a normal mean. From 1890 to 1893 Edgeworth, reacting to work by Galton, gave the first fully developed mathematical examination of correlation and its relation to the multivariate normal distribution (Edgeworth 1892a, b). Edgeworth showed how the constants of a multivariate normal distribution could be expressed in terms of pairwise correlation coefficients (and hence how the conditional expectation of one variable given others could be expressed in terms of correlation coefficients), and he investigated how a correlation coefficient could be estimated from data. His work gave what may be the earliest version of what has come to be called Pearson's product moment estimate (or Pearson's *r*). Incidentally, it was Edgeworth who coined the term 'coefficient of correlation', as Galton had used 'index'.

Edgeworth's work on correlation had an immense influence upon Karl Pearson, and through him upon all 20th-century work on this topic. In the 1890s, Edgeworth's statistical work became increasingly occupied by a competition with Karl Pearson as to who could best model skew data. Pearson, with his family of skew curves that included gamma distributions and a scheme (the method of moments) for selecting a curve within this family, is generally conceded to have won the contest. Edgeworth at one time or another tried three different approaches. One of these (the 'method of translation', or fitting a normal curve to transformed data) has become popular in more recent times. Another (fitting separate half-normal curves to the left and right sides of the distribution) has been largely forgotten. The third was based upon what we now call the Edgeworth Series. The essence of Edgeworth's approach was to generalize the central limit theorem by the inclusion of correction terms, terms that appeared in the derivation of the distribution of sums but which became negligible if the number of terms in the sum was large. The idea was that skew distributions found in nature

were skew because they were aggregates of relatively small numbers of non-normal components. Edgeworth was thus taking a theoretical approach, one that he felt was more appealing than Pearson's more ad hoc approach. The Edgeworth Series was foreshadowed in his work as early as 1883 (when he found it as a series solution to the heat equation), but the full development came later (Edgeworth 1905), and the labour he put into it after 1895 was immense, and largely unrewarded. His attempts to provide a methodology for fitting the series to data attracted few followers, Arthur Bowley being the only important one. Bowley's brave attempt to explain the method in his assessment of Edgeworth's work (Bowley 1928) was only marginally more readable than Edgeworth's own many efforts on this. Ironically, later statisticians (notably Harold Cramér, see Cramér 1972) have found that Edgeworth's mode of arranging correction terms was far superior to alternatives proposed by Bruns, Gram and Charlier, and the Edgeworth Series has become an important technique for approximating sampling distributions (rather than data distributions, as Edgeworth had intended).

In addition to these major themes, Edgeworth's work abounds in minor nuggets. The largest of these may be a series of papers in 1908–9 that we can now recognize as containing the germ of a proof of the asymptotic efficiency of maximum likelihood estimates. In a contentious 1935 meeting of the Royal Statistical Society this work was pointed out to R.A. Fisher by Bowley as an unacknowledged predecessor, although it seems doubtful that it had any influence on Fisher (see Pratt 1976). Of more importance was Edgeworth's work on index numbers and on the theory of banking. While his work on index numbers is more properly treated with his economic work, it is worth noting here that he was a pioneer in the application of probability to the analysis and choice of index numbers. In regard to banking, based upon statistical considerations, he promulgated in 1888 the rule that the reserves of a bank need only be proportional to the square root of its liabilities (Edgeworth 1888).

In all Edgeworth's work one is constantly coming upon minor, often paradoxical observations (see for example, Stigler 1980) that reveal the depth of his understanding, the subtlety of his thoughts, and a grasp of mathematics that seems quite at odds with his lack of formal training in the subject. Edgeworth was an independent thinker upon statistical matters, though he was perhaps the earliest to appreciate and follow up on Galton's innovative concepts of regression and correlation. Edgeworth's most important influence was upon Karl Pearson, though Pearson was chary in his recognition of this influence. Taken together, Galton, Edgeworth and Pearson shaped modern statistics to a greater degree than any other individual or group before R.A. Fisher. Edgeworth's works on statistics number at least 75, and it is rare to find one that is self-contained. Bowley (1928) made an attempt to summarize all of Edgeworth's statistical work, and he gave a bibliography of most of it. Stigler (1978, 1986) gives a more recent assessment, and comments upon different aspects of Edgeworth's work can be found in papers by Kendall (1968, 1969) and Pratt (1976).

Bibliography

Bowley, A.L. 1928. *F.Y. Edgeworth's contributions to mathematical statistics.* London: Royal Statistical Society. Reprinted, New York: Augustus M. Kelley, 1972.

Cramér, H. 1972. On the history of certain expansions used in mathematical statistics. *Biometrika* 59: 205–207.

Edgeworth, F.Y. 1885a. Observations and statistics. An essay on the theory of errors of observation and the first principles of statistics. *Transactions of the Cambridge Philosophical Society* 14: 138–169.

———. 1885b. Methods of statistics. *Jubilee Volume of the Statistical Society*: 181–217.

———. 1885c. On methods of ascertaining variations in the rate of births, deaths, and marriages. *Journal of the Royal Statistical Society* 48: 628–649.

———. 1888. The mathematical theory of banking. *Journal of the Royal Statistical Society* 51: 113–127.

———. 1892a. Correlated averages. *Philosophical Magazine* 34(Fifth Series): 190–204.

———. 1892b. The law of error and correlated averages. *Philosophical Magazine* 34(Fifth Series): 429–438, 518–26

———. 1905. The law of error. *Transactions of the Cambridge Philosophical Society* 20: 36–65, 113–41

———. 1908. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society* 71: 381–397, 499–512, 651–678; 72, 81–90
Kendall, M.G. 1968. Francis Ysidro Edgeworth, 1845–1926. *Biometrika* 55: 269–275.
———. 1969. The early history of index numbers. *Review of the International Statistical Institute* 37: 1–12.
Pratt, J. 1976. F.Y. Edgeworth and R.A. Fisher on the efficiency of maximum likelihood estimation. *Annals of Statistics* 4: 501–514.
Stigler, S.M. 1978. Francis Ysidro Edgeworth, statistician (with discussion). *Journal of the Royal Statistical Society, Series A* 141: 287–322.
———. 1980. An Edgeworth curiosum. *Annals of Statistics* 8: 931–934.
———. 1986. *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Belknap Press of the Harvard University Press.

Edgeworth Price Cycles

Michael D. Noel

Abstract

Edgeworth price cycles refer to an asymmetric pattern of prices that result from a dynamic pricing equilibrium among competing oligopolists. The resulting time series takes on a sawtooth shape – many small price decreases interrupted only by occasional large price increases. Maskin and Tirole (Econometrica 56(3): 571–599, 1988) formalized the theory, and later extensions were provided by Eckert (International Journal of Industrial Organization 21(3): 151–170, 2003) and Noel (Journal of Economics and Management Strategy 17(2): 345–377, 2008). Edgeworth price cycles are the leading theory for explaining the asymmetric price cycles that appear in many US, Canadian, Australian and European retail gasoline markets (e.g. Noel (Review of Economics and Statistics 89(2): 324–334, 2007a), Eckert (Canadian Journal of Economics 35(1): 52–77, 2002), Doyle et al. (Energy Economics 32(3): 651–660, 2010), Wang (Journal of Political Economy 117(6): 987–1030, 2009b)). While the gasoline cycles continue to generate public concern with claims of collusion often raised, the current evidence favours Edgeworth price cycles being the result of stronger competition and the source of lower retail gasoline prices.

Keywords

Markov strategies; Markov perfect equilibria; Cournot model; Retail petrol markets

JEL Classifications

L13; L81; L92

Introduction

Edgeworth price cycles refer to an asymmetric pattern of prices that is generated by a dynamic pricing equilibrium among competing oligopolists under certain simple assumptions. Most notably, the oligopolists are assumed to compete in prices, follow Markov strategies, and face relatively high price elasticities for their good. The time series of market prices under this equilibrium takes on a sawtooth shape, with many small price decreases interrupted only by occasional large price increases. The asymmetric price pattern is repeated over and over, even in the absence of any supply or demand shocks.

Edgeworth price cycles are the leading theory behind the asymmetric price cycles that appear in many retail gasoline markets around the world. First observed in some US cities in the 1960s, they have become commonplace in many US, Canadian, Australian and European retail gasoline markets (e.g. Noel (2007a), Eckert (2002), Doyle et al. (2010), Wang (2009b)) and visually are very similar to the theoretical cycles. A single cycle is often a week or two long with amplitude up to about 10% of the price.

Two waves of literature examine these cycles empirically. The first investigates the cause of the retail gasoline price cycles. The near consensus in the literature is that the cycles are Edgeworth price cycles. The location and shape of the cycles and the behaviour of different types of firms along the cycles support the Edgeworth price cycles theory.

The second wave of literature examines the welfare effects of the cycles relative to a stable price equilibrium. The literature is young, but the results currently favour the conclusion that the price cycles are indicative of stronger competition and the source of lower prices for consumers.

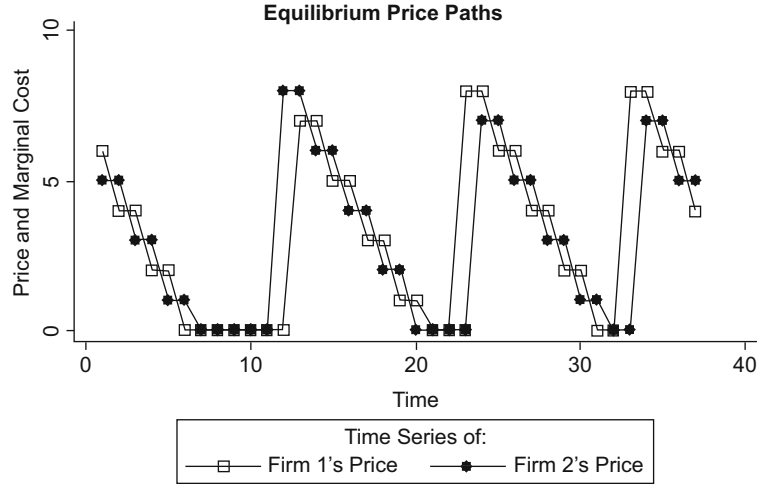
The Theory of Edgeworth Price Cycles

The notion of a competitively driven, dynamic, asymmetric price cycle dates back to Edgeworth (1925). Edgeworth was a strong critic of the Cournot model and argued that when marginal costs were increasing (or firms were capacity constrained in the extreme case), prices in oligopolistic competition would not be stable as in Cournot's model. Instead, they would change continually along an asymmetric price cycle.

Firms would undercut one another to gain market share until prices were low enough that one firm could profitably raise the price and serve the residual demand left over from the capacity constrained firm. Interestingly, Edgeworth considered his cycle a disequilibrium, as the notion of equilibrium was then equated to stable prices.

The seminal theory paper on Edgeworth price cycles is by Maskin and Tirole (1988), who gave the cycles their name. Maskin and Tirole assume two identical and infinitely lived firms with high discount factors that sell homogeneous goods and compete in prices. Firms are restricted to using Markov strategies, which in this context means that a firm's pricing decision depends only on the price of the other firm currently in effect. Demand is constant and costs are zero. Maskin and Tirole show that in this setting two possible types of Markov perfect equilibria could result. The first generates stable prices over time, while the second results in asymmetric price cycles – that is, Edgeworth price cycles – in equilibrium.

Figure 1 shows firm prices in a cycling equilibrium. The mechanism operates as follows. Starting from prices relatively high above marginal cost, firms alternately and repeatedly undercut one another's price by the smallest possible amount. Because the goods are identical, this is sufficient to fully steal total market demand. Undercutting continues until prices fall all the way to marginal cost (zero in this example). At that point, there is no gain from lowering prices further, but there is a gain from raising them. If one firm were to raise its price to a much higher level, the other firm would surely respond with a higher price too, just slightly undercutting that of the first firm. As a result, when prices are at marginal cost, firms play a war of attrition, each mixing between a higher price and maintaining the marginal cost price. Eventually, one firm relents by restoring its price to a high level, the other follows, and a new round of undercutting



Edgeworth Price Cycles, Fig. 1 Theoretical Edgeworth cycle

begins. The cycle repeats over and over even in the absence of cost and demand shocks.

The theory of Edgeworth price cycles is the leading theory for explaining the asymmetric price cycles found in retail gasoline markets. However, not all of its assumptions are well suited to retail gasoline. Retail gasoline is not perfectly homogeneous, marginal costs are not constant, there are more than two firms, and so on. To address this, Noel (2008) extends the model to allow for fluctuating costs, capacity constraints, product differentiation, triopoly situations, and other extensions. Noel shows that Edgeworth price cycles are robust when product differentiation or capacity constraints are not too strong. Noel also shows that cycles are robust to triopoly, but now with challenges in the form of false starts. These occur when the first firm to increase price abandons its high price altogether, after waiting too long for others to follow. Since false starts make it more costly for a firm to be the first to raise its price, the cycle peak and trough prices move lower and average prices fall from increased competition. In another important extension, Eckert (2003) showed that if firms share the market very unequally at equal prices (interpretable as the case of asymmetrically sized firms), Edgeworth price cycles are more likely.

Empirical Evidence of Edgeworth Price Cycles

An early criticism of the theory, dating back to at least Nichol (1935), was that the model failed to predict the experience of any known real-world markets. There has also been very limited success in generating Edgeworth cycles in laboratory experiments (Leufkens and Peeters 2008).

This all changed when – with the availability of new high-frequency and station-specific datasets – asymmetric retail price cycles were discovered in gasoline markets in many countries in the 1990s and 2000s. As of 2010, cycles have been observed in many retail gasoline markets in the USA, Canada, Australia, Norway, Germany and Belgium, with other discoveries sure to come. (Cycles had been detected in a few US cities in the late 1960s and early 1970s, but they received little attention from economists at the time and then disappeared from view for thirty more years.) Recently, inverted asymmetric price cycles have also been found in keyword advertising auctions at leading search engines (inverted because competition is among buyers instead of sellers) (Zhang 2005).

The first generation of studies in the literature sought a cause for the empirical asymmetric cycles, and the results of that literature strongly point to Edgeworth price cycles (to name a few: Castanias and Johnson (1993), Lewis and Noel (in press), Lewis (2009a, b), Doyle et al. (2010), in US retail gasoline markets; Noel (2007a, b, 2009, 2010a), Eckert (2002, 2003), Eckert and West (2004), Atkinson (2009) in Canada; Wang (2009a, b) in Australia; and Zhang (2005) in Internet auction markets). One opposing view is that of Foros and Steen (2008), who examine Norway and argue that the cycles there, although similar to those in other countries, are possibly a form of pure collusion.

The first known publication that noted asymmetric price cycles in retail gasoline was Allvine and Patterson (1974), who showed cycles in several US cities in the late 1960s and early 1970s. Castanias and Johnson (1993) reported summary statistics on the cycles in Los Angeles area and noted the resemblance between those cycles and the then newly published Maskin and Tirole (1988) theory article. In Los Angeles, a single cycle lasted about one to two months with an amplitude of roughly 25% of the price.

The first full-length scientific papers were written about Edgeworth cycles in Canada, where most large cities experienced price cycles beginning from at least the late 1980s. The cycles ranged from weekly up to bimonthly in duration and longer. Eckert (2003) showed that price volatility in retail gasoline prices in Canadian cities were consistent with the general predictions of his extension of the Edgeworth cycle model.

Noel (2007a) specifically modelled the asymmetric price movements in Canadian cities and found that a greater market share of aggressive independent firms resulted in cycles that were *faster*, *taller* and *less* asymmetric, consistent

with the theory. Noel (2007b) examined station-specific data in Toronto, Canada, and showed that behaviours of differently sized firms were also consistent with the theory. Large refiner-retailers tended to lead price increases, and smaller independents tended to lead price decreases. The duration of cycles in Toronto averaged a week and the amplitude was about 8% of the price. Atkinson (2009) found similar results with high-frequency price data in Guelph, Canada.

Retail gasoline price cycles were rediscovered in dozens of US cities in the early 2000s, although it is uncertain how long they existed prior to that. The cycles were weekly in most cities but biweekly in a few. Lewis (2009a, b) support that the cycles in the Midwest US are Edgeworth price cycles. The amplitude of the cycles averaged 7% of the retail price. Doyle et al. (2010) present further evidence that cycles were more likely with more large, price-aggressive independents, consistent with their extension of the theory.

In major Australian cities and many smaller ones, retail gasoline price cycles were weekly in duration, except in Perth where they were sometimes biweekly. Wang (2009a) estimates especially high cross price elasticities in Perth, consistent with the theory, and Wang (2009b) shows that the pattern of price increases among the leading firms is consistent with the use of mixed strategies, as predicted in a symmetric Edgeworth price cycles model.

Foros and Steen (2008) examine the cycles in Norway which share the characteristics of cycles elsewhere. They offer the opposing view that the Norwegian cycles may be a pure collusion story on the basis that the large price increases occur in a short window on late mornings and early afternoons on Mondays. While there is no direct evidence of widespread collusion in Norway or elsewhere, isolated instances of individual dealers colluding with one another can occur in cycling markets, just as they can in non-cycling markets (see Wang (2008) for the case of Ballarat, Australia, and Erutku and Hildebrand (2010) for the case of four towns in Quebec, Canada).

Edgeworth price cycles have not been well understood in many circles and the large sweeping price increases understandably raise antitrust scrutiny. The second wave of the Edgeworth price cycles empirical literature examines the impact that cycles have on price and welfare.

Most important and difficult is the question of how markups rise and fall with the presence of Edgeworth cycles. Noel (2002) shows that in Canada markups within the same city in a nearby time period (i.e. controlling for time–city effects) were one cent lower just after cycles began compared with before. The consumer gain of cycling in his sample cities is CDN$48 million per year. Doyle et al. (2010) show that in the USA, markups in cycling cities are one to two US cents per gallon lower than in non-cycling cities, controlling for market structure and other observables. Wang (2009b), however, finds that prices were 1.8 cents lower when cycles temporarily ceased in the four months after the passing of the 24 hour price change pre-notification law in Perth, Australia. Noel (2010b) notes it would not be unusual for prices to temporarily fall in such a situation, as firms jostle for position and generate a string of false starts as they adjust to the new notification requirements.

Lewis (2009a) and Lewis and Noel (in press) argue for another pro-competitive aspect to Edgeworth price cycles – that they more effectively anchor prices to wholesale costs over the long run. In non-cycling markets, the well-known rockets and feathers phenomenon is that prices rise quickly after a cost increase but fall slowly after a cost decrease (Borenstein et al. 1997). Noel (2009) and Eckert (2003) even show that in cycling markets, the presence of the asymmetric cycle can generate or magnify such an effect. However, compared to non-cycling markets, in cycling markets the cycles are effective in returning prices to costs at every trough. Lewis (2009a) shows that retail gasoline prices in US markets with Edgeworth price cycles fell much more quickly in the months after Hurricane Katrina than those cities without, for a relative gain of US$1.33 million per 100,000 people. Lewis and Noel (in press) look more comprehensively at 72 US cities and show that cost shocks are passed through two to three times faster in cities with Edgeworth price cycles than in cities

without, substantially reducing the rockets and feathers welfare loss relative to non-cycling cities.

Noel (2010a) argues for another hidden benefit of Edgeworth price cycles. Very simple purchase timing strategies can allow price elastic consumers in Toronto to easily reduce their gasoline expenditures by 4% relative to purchasing at random times. The gain would be even greater in other markets (e.g. Norway or Australia) where cycle troughs are even more easily predictable. Firm price reoptimization in response to large numbers of consumers timing the cycles could limit the gain, of course, but as the Australian experience shows, the cycles remain active and strong even when there is significant consumer awareness of them.

The greater implication of Noel (2010a) is that to the extent that consumers use purchase timing strategies that are not observed by the researcher, the benefit of Edgeworth price cycles is understated. This is because when consumers can time purchases to periods of lower prices during cycles, the more economically relevant measure of *quantity-weighted* prices is likely to be lower than the unweighted average price in markets with cycles, but the same in markets without. A comparison of average quantity-weighted prices under periods of cycling and non-cycling would then reveal a greater price advantage to consumers in markets with Edgeworth price cycles.

Conclusion

Edgeworth price cycles are asymmetric price cycles generated from equilibrium behaviour in a game of oligopolistic price competition. Firms repeatedly steal market demand from one another by undercutting down to marginal cost. Firms then sequentially increase prices back to the top of the cycle and begin undercutting again.

The empirical literature strongly favors the asymmetric price cycles in retail gasoline markets being generated by an Edgeworth price cycles process. While research continues, the weight of the current evidence also points to the conclusion that Edgeworth price cycles are indicative of stronger competition. They benefit consumers with lower and more efficient prices relative to the less controversial stable price equilibrium.

The literature on Edgeworth price cycles continues to grow. An obvious direction for future work is to search for and uncover additional examples of Edgeworth price cycles outside of retail gasoline. Further extensions to the theory can help guide this search, and conversely the findings of the search can suggest new extensions to the theory to help identify the factors most critical to cycle generation. Finally, the welfare effects of Edgeworth price cycles have only recently begun to be understood. An important direction of future research would be to study and quantify these effects further, with obvious and important antitrust and policy implications.

See Also

- ▶ Cournot Competition
- ▶ Edgeworth, Francis Ysidro (1845–1926)
- ▶ Gasoline Markets
- ▶ Markov Equilibria in Macroeconomics
- ▶ Markov Processes

Bibliography

Allvine, F., and J. Patterson. 1974. *Highway robbery: An analysis of the gasoline crisis*. Bloomington: Indiana University Press.

Atkinson, B. 2009. Retail gasoline price cycles: Evidence from Guelph, Ontario using bi-hourly station-specific retail price data. *Energy Journal* 30(1): 85–110.

Borenstein, S., A. Cameron, and R. Gilbert. 1997. Do gasoline markets respond asymmetrically to crude oil price change. *Quarterly Journal of Economics* 112: 305–339.

Castanias, R., and H. Johnson. 1993. Gas wars: Retail gasoline price fluctuations. *Review of Economics and Statistics* 75(1): 171–174.

Doyle, J., E. Muehlegger, and K. Samphantharak. 2010. Edgeworth cycles revisited. *Energy Economics* 32(3): 651–660.

Eckert, A. 2002. Retail price cycles and response asymmetry. *Canadian Journal of Economics* 35(1): 52–77.

Eckert, A. 2003. Retail price cycles and presence of small firms. *International Journal of Industrial Organization* 21(3): 151–170.

E

Eckert, A., and D. West. 2004. Retail gasoline price cycles across spatially dispersed gasoline stations. *Journal of Law and Economics* 22: 997–1015.
Edgeworth, F.Y. 1925. The pure theory of monopoly. In *Papers relating to political economy*, vol. I, ed. F.Y. Edgeworth, 111–142. London: Macmillan.
Erutku, C., and V. Hildebrand. 2010. Conspiracy at the pump. *Journal of Law and Economics* 53(1): 223–237.
Foros, O., and F. Steen. 2008. Gasoline prices jump up on Mondays? An outcome of aggressive competition? *CEPR Working Paper DP6783*.
Leufkens, K., and R. Peeters. 2008. Focal prices and price cycles in an alternating price duopoly experiment. *METEOR Working Paper RM/08/021*.
Lewis, M. 2009a. Temporary wholesale gasoline price spikes have long lasting retail effects: The aftermath of Hurricane Rita. *Journal of Law and Economics* 52(3): 581–606.
Lewis, M. 2009b. Price leadership and coordination in retail gasoline markets with price cycles. *The Ohio State University Working Paper*.
Lewis, M., and M. Noel. In press. The speed of gasoline price response in markets with and without Edgeworth cycles. *Review of Economics and Statistics*.
Maskin, E., and J. Tirole. 1988. A theory of dynamic oligopoly II: Price competition, kinked demand curves and Edgeworth cycles. *Econometrica* 56(3): 571–599.
Noel, M. 2002. *Edgeworth price cycles in retail gasoline markets*. PhD dissertation, MIT.
Noel, M. 2007a. Edgeworth price cycles, cost-based pricing and sticky pricing in retail gasoline retail markets. *Review of Economics and Statistics* 89(2): 324–334.
Noel, M. 2007b. Edgeworth price cycles: Evidence from the Toronto retail gasoline market. *Journal of Industrial Economics* 55(1): 69–92.
Noel, M. 2008. Edgeworth price cycles and focal prices: Computational dynamic Markov equilibria. *Journal of Economics and Management Strategy* 17(2): 345–377.
Noel, M. 2009. Do gasoline prices respond asymmetrically to cost shocks? The effect of Edgeworth cycles. *RAND Journal of Economics* 40(3): 582–595.
Noel, M. 2010a. Edgeworth cycles and intertemporal price discrimination. *UCSD Working Paper*.
Noel, M. 2010b. Edgeworth cycles, competition, and antitrust. *UCSD Working Paper*.
Wang, Z. 2008. Collusive communication and pricing coordination in a retail gasoline market. *Review of Industrial Organization* 32(1): 35–52.
Wang, Z. 2009a. Station level gasoline demand in an Australian market with regular price cycles. *Australian Journal of Agricultural and Resources Economics* 53: 467–483.
Wang, Z. 2009b. Mixed strategies in oligopoly pricing: Evidence from gasoline price cycles before and under a timing regulation. *Journal of Political Economy* 117(6): 987–1030.
Zhang, M. 2005. Finding Edgeworth cycles in online advertising auctions. *MIT mimeo*.

Edgeworth, Francis Ysidro (1845–1926)

John Creedy

Abstract

Edgeworth was a major figure in the development of neoclassical economics, and one of its most original theorists, making a wide range of lasting contributions. After describing his approach to economics, this article discusses his early work in moral philosophy, which had a strong influence on his economics. His important contribution to the theory of exchange, focusing on indeterminacy and the role of the number of traders, is examined. His later work on monopoly, international trade and taxation are then briefly discussed.

Keywords

Arbitration; Assumptions; Barter; Bentham, J.; Bickerdike, C.; Bilateral monopoly; Biological analogies; Bootstrap; Butler, J.; Calculus of variations; Coalitions; Collusion; Combinations; Competitive (price-taking) equilibrium; Complements: *see* substitutes and complements; Conjectural variations; Contract curve; Correlation coefficient; Cournot, A.; Darwin, C.; Deductive method; Determinacy and indeterminacy of exchange; Distributive justice; Duopoly; Edgeworth box; Edgeworth, F.; Egoism; Equilibrium in exchange; Experimental psychology; First fundamental theorem of welfare economics; Giffen good; Harsanyi, J.; Historical School; Hotelling, H.; Idealism; Immiserizing growth; Indeterminacy of contract; Indifference map; Inference; International trade; International values, theory of; Intuitionism; Jevons, W.; Keynes, J. M.; Lagrange multipliers; Laplace, P.; Launhardt, C.; Law of indifference; Marshall, A.; Mathematical economics; Mechanical analogies; Mill, J. S.; Monotonicity; Moral philosophy; Negative income tax; Neoclassical; No-profit

entrepreneur; Offer curve; Optimal distribution; Optimal tariffs; Paley A. and M.; *Palgrave's Dictionary of Political Economy*; Partial equilibrium theory; Pearson distributions; Physical sciences; Pigou, A.; Probability; Progressive and regressive taxation; Rate of exchange; Reciprocal demand curve; Recontracting; Royal Statistical Society; Rules of conduct; Sacrifice theory of tax incidence; Saddle point; Schumpeter, J.; Sidgwick, H.; Social contract; Social welfare function; Statistical inference; Substitutes and complements; Tax incidence; Taxation, theory of; Transformations; Utilitarianism; Utility functions; Utility maximization; Vickrey, W.; Walras, L.

JEL Classifications
B31

Biographia

Francis Ysidro Edgeworth (1845–1926) was born in Edgeworthstown in County Longford, Ireland. The background into which he was born was dominated by the 'larger than life' figure of his grandfather Richard Lovell Edgeworth (1744–1817), whose life was documented in a two-volume memoir (1820) by his oldest daughter, the famous novelist Maria Edgeworth (1767–1849). Richard Lovell's many scientific and mechanical experiments were helped by his strong association with the Lunar Society of Birmingham, whose members included Watt, Bolton, Wedgwood, Priestley, Darwin and Galton. In addition, Maria's scientific acquaintances included Davy, Humboldt, Herschel, Babbage, Hooker and Faraday. The marriage of F. Y. Edgeworth's cousin Harriet Jessie Edgeworth (daughter of Richard Lovell's seventh and youngest son Michael Pakenham, 1812–1881) to Arthur Gray Butler provided links with another large and eminent academic family. These connections extend even further since A. G. Butler's sister, Louisa Butler, married Francis Galton, a cousin of Charles Darwin.

Richard Lovell's sixth son, and 17th surviving child, was Francis Beaufort Edgeworth (1809–1846), who met his wife, Rosa Florentina Eroles, the daughter of a Spanish refugee from Catalonia and then aged 16, while on the way to Germany to study philosophy; they married within three weeks in 1831. F. Y. Edgeworth was their fifth son. With his family background and his knowledge of French, German, Spanish and Italian, Edgeworth had wide international sympathies. On the family background, see Butler and Butler (1927) and for a full-length treatment of Edgeworth's work, see Creedy (1986).

Edgeworth was educated by tutors in Edgeworthstown until the age of 17, when in 1862 he entered Trinity College Dublin to study languages. In 1867 Edgeworth entered Exeter College, Oxford, but after one term transferred to Magdalen Hall. He transferred to Balliol in 1868, where in Michaelmas 1869 he obtained a first in Literae Humaniares. He was called to the bar in 1877, the same year in which his first book, *New and Old Methods of Ethics,* was published. Edgeworth applied unsuccessfully for a professorship of Greek at Bedford College, London, in 1875, but later lectured there on English language and literature for a brief period from late 1877 to mid-1878. He had earlier lectured on logic, mental and moral sciences and metaphysics to prospective Indian civil servants, at a private institution run by a Mr. Walter Wren. In 1880 he applied for a chair of philosophy, also unsuccessfully, but began lecturing on logic to evening classes at King's College London. Soon after the publication of his second book, *Mathematical Psychics,* in 1881, he applied for a professorship of logic, mental and moral philosophy and political economy at Liverpool. Testimonials for two of Edgeworth's applications were given by Jevons (see Black 1977, V, pp. 98, 145) and Marshall.

Edgeworth had to wait until 1890 until he obtained a professorial appointment: this was at King's College London, where he succeeded Thorold Rogers in the Tooke Chair of Economic Science and Statistics. In the next year, 1891, he again succeeded Rogers, this time to become Drummond Professor and Fellow of All Souls' College, Oxford, a position he held until his

retirement in 1922. Edgeworth therefore finally settled in Oxford at the age of 46 in what was to become one of the most illustrious British chairs in economics. At the same time he became the first editor of the *Economic Journal.* He was editor or co-editor from its first issue until his death. He was supported by Henry Higgs from 1892 to 1905, when the latter became the Prime Minister's Private Secretary, with further assistance provided at a later stage by Alfred Hoare. Keynes was a co-editor for 15 years. After a tremendously creative period of the late 1870s and 1880s, Edgeworth had become firmly established as the leading economist, after Marshall, in Britain.

In addition to his work in economics, Edgeworth began a series of statistical papers in 1883. He was President of section F of the British Association in 1889, a position he held again in 1922. Edgeworth's work on mathematical statistics played an increasingly important role. Indeed, of about 170 papers which he published, approximately three-quarters were concerned with statistical theory. He became a Guy Medalist (Gold) of the Royal Statistical Society in 1907 and was President of the Society during 1912–1914. His main contributions to statistics concern work on inference and the law of error, the correlation coefficient, transformations (what he called 'methods of translation'), and the 'Edgeworth expansion'. The latter, a series expansion which provides an alternative to the Pearson family of distributions, has been widely used (particularly since the work of Sargan 1976) to improve on the central limit theorem in approximating sampling distributions. It has also been used to provide support for the bootstrap in providing an Edgeworth correction. Edgeworth's work in probability and statistics has been collected by McCann (1996). His third and final book was *Metretike*: *or the Method of Measuring Probability and Utility* (1887). These contributions are not examined here; see Bowley (1928) and Stigler (1978).

Approach to Economics

A dominant characteristic of Edgeworth's approach to economics is that it is mathematical, characterized by an original use of techniques, although he does not appear to have received a formal training in mathematics. However, he came to economics from moral philosophy. The central question of distributive justice, rather than simply the application of mathematics, dominated his attitude towards economics. His main argument was that mathematics provided powerful assistance to 'unaided' reason, and could check the conclusions reached by other methods. Thus:

> He that will not verify his conclusions as far as possible by mathematics, as it were bringing the ingots of common sense to be assayed and coined at the mint of the sovereign science, will hardly realise the full value of what he holds, will want a measure of what it will be worth in however slightly altered circumstances, a means of conveying and making it current. (1881, p. 3)

Edgeworth's approach contrasts sharply with that of Marshall. The contrast between Edgeworth and Marshall was neatly summarized by Pigou as follows:

> During some thirty years until their recent deaths in honoured age, the two outstanding names in English economics were Marshall ... and Edgeworth... Edgeworth, the tool-maker, gloried in his tools . . . Marshall, on the other hand, had what almost amounted to an obsession for hiding his tools away. (Pigou and Robertson, 1931, p. 3)

Although both men turned to economics from mathematics and moral philosophy, Marshall generally used biological analogies, and was concerned with developing maxims. In contrast, Edgeworth generally used mechanical analogies, and was more concerned with developing theorems.

In the 1880s and 1890s the deductive method encountered a great deal of criticism, especially from the 'Historical School' of economists. Edgeworth's defence of the deductive method often involved showing how other economists had advocated its use. His interest in the natural sciences often led him to make comparisons with scientific laws, and especially to show that the physical sciences also relied on abstraction and approximation.

Edgeworth argued carefully that the assumptions used in economics are often untestable, and he therefore took precautions against the

accusation of 'plucking assumptions from the air'. He was conscious of the fact that the difficulty is in making the crucial abstractions which make the particular problem under consideration tractable, but which are not question begging. His attitude to many a priori assumptions was directly related to his approach to statistical inference. In *Mathematical Psychics,* for example, he referred to 'the first principle of probabilities, according to which cases about which we are equally undecided ... count as equal' (1881, p. 99). This was then transferred to economics. The appropriate assumption was that all feasible values, say, of elasticities, were equally likely, until evidence is obtained. Hence, 'There is required, I think ... in order to override the a priori probability, either very definite specific evidence, or the consensus of high authorities' (1925, ii, pp. 390–391). This also illustrates Edgeworth's attitude to authority and his many allusions to the views of other leading economists. Price (1946, p. 38) referred to his frequent 'reference to authority for . . . support of tentative opinion waveringly advanced'.

Edgeworth was also prone to stress negative results. For example, in discussing taxation, where the criterion of minimum sacrifice does not alone provide a simple tax formula, he stated:

> Yet the premises, however inadequate to the deduction of a definite formula, may suffice for a certain negative conclusion. The ground which will not serve as the foundation of the elaborate edifice designed may yet be solid enough to support a battering-ram capable of being directed against simpler edifices in the neighbourhood. (1925, p. 261)

Edgeworth's position as editor of the *Economic Journal* enabled him to combine both his critical attitude and his appetite for a wide range of reading. He contributed 32 book reviews, and in sending books to other reviewers he would include 'apposite remarks on particular points in the text' (Bowley 1934, p. 123). These reviews should also be placed beside his 17 reviews in the Academy, and 131 articles in the original *Palgrave's Dictionary of Political Economy.* Furthermore, Edgeworth's later articles in the *Economic Journal,* such as those on international trade and on taxation, took the form of extended commentaries on contemporary work.

Early Work in Moral Philosophy

Before turning to economics, Edgeworth published a brief note in *Mind* in 1876, and his first (privately printed) book on New and Old Methods of Ethics in 1877. The description by Keynes of Edgeworth's first book could just as well be applied to his other two books:

> Edgeworth's peculiarities of style, his brilliance of phrasing, his obscurity of connection, his inconclusiveness of aim, his restlessness of direction, his courtesy, his caution, his shrewdness, his wit, his subtlety, his learning, his reserve –all are there full-grown. Quotations from the Greek tread on the heels of the differential calculus. (Keynes, 1972, p. 257)

The main focus of this early work, strongly influenced by the great Cambridge philosopher Henry Sidgwick (1838–1900), was to examine in detail the implications of utilitarianism for the optimal distribution of resources. Edgeworth's special and original contribution was to apply advanced mathematics to this problem. Edgeworth's approach was dominated by his utilitarianism, but the influence of contemporary psychological research and the impact of evolutionary ideas can also be traced. Both aspects led to explicit consideration of differences between individuals and changes which take place over time.

Edgeworth was also influenced by the major fierce debates in the last half of the 19th century between egoism, evolutionism, idealism, intuitionism, and of course utilitarianism. His brand of utilitarianism became extremely eclectic, and embraced the majority of the above principles (except for those of the Hegelian idealists) while regarding utilitarianism as the 'sovereign principle'. His note in *Mind* discussed Matthew Arnold's views of Joseph Butler, who had examined egoism at great length. Arnold had argued that Butler's term 'self love' should be interpreted to mean 'the pursuit of our temporal good'. However, Edgeworth argued that egoism and utilitarianism could be subsumed under the same principle. He believed Butler to be saying, 'duty and interest are perfectly coincident; for the most part in this world, but entirely and in every

instance, if we take in the future and the whole' (1876, p. 571).

Edgeworth generally distinguished between 'impure' and 'pure' utilitarianism. In the latter case individuals are assumed to be concerned with the welfare of society as a whole. The former case in fact corresponds more closely with a 'short term' version of egoism. Economic exchange can usefully be analysed in terms of 'jostling egoists', but he believed that ultimately individuals would evolve to become pure utilitarians. A reason for believing that individuals would make such a transition was later to be developed by Edgeworth in the form of his contractarian justification of utilitarianism as the appropriate principle of distributive justice.

Edgeworth's early utilitarianism was influenced by his wide knowledge of work in experimental psychology. In his books of 1877 and 1881 there are many references to the work of Delboeuf, Fechner, Helmholtz, Weber and Wundt. These references occur in the context of discussing the nature of utility functions and, although Edgeworth at this time was not aware of the earlier work of Jevons, the same range of psychological work was also important to Jevons. Edgeworth in 1877 explicitly suggested, in connection with Fechner, that an additive form would not be appropriate.

A further aspect, of Edgeworth's utilitarianism is his attitude towards authority. An important issue for early utilitarians involved the nature of inductive evidence about the consequences of acts. Most people cannot know the full consequences of their acts, so that rules of moral conduct must be followed (in contrast with intuitionism where individuals are assumed to have immediate consciousness of moral rules). In arriving at such rules, the opinions of highly regarded individuals are taken to be credible though it may not be possible to show conclusively that they are 'correct'. Edgeworth argued, for example, that 'we ought to defer even to the undemonstrated dicta and opinions of the wise, who have a power of mental vision acquired by experience' (1925, ii, p. 149).

Edgeworth defined the problem of determining the optimal utilitarian distribution as follows: 'given a certain quantity of stimulus to be distributed among a given set of sentients . . . to find the law of distribution productive of the greatest quantity of pleasure' (1877, p. 43). In treating this problem mathematically Edgeworth used Lagrange multipliers, without any explanation, and concluded that, 'unto him that hath greater capacity for pleasure shall be added more of the means of pleasure' (1877, p. 43). In using Lagrange multipliers Edgeworth was also careful to discuss possible complications, referring to the possibility of multiple solutions and explicitly discussing corner solutions and inequality constraints.

Further complexities were then examined, where Edgeworth emphasized that utilitarianism implies equality of the 'means of pleasure' only under a special set of assumptions, and in the general case the prescribed solution will be some form of inequality. In dealing with the distribution of effort, he argued not surprisingly that most work should be provided by those most capable of providing it. In a yet more general treatment of the problem, Edgeworth used the calculus of variations, but again provided the reader with virtually no help in following his mathematical argument. Edgeworth's analysis of the utilitarian optimal distribution was continued in his paper on 'The Hedonical Calculus' (1879), which was later reprinted as the third part of *Mathematical Psychics.*

Early Work in Economics

The turning point in Edgeworth's work was his introduction to Jevons in 1879 by a mutual friend James Sully, who in 1878 moved to Hampstead, where Edgeworth had lodgings in Mount Vernon and where Jevons also lived; see Sully (1918, pp. 180, 223). His first knowledge of Marshall came from Jevons, who 'highly praised the then recently published Economics of Industry' (in Pigou 1925, p. 66). Edgeworth became interested in the problem of the indeterminacy of the rate of exchange, arising from the existence of only a small number of transactors. This led rapidly to Edgeworth's second and most important

book *Mathematical Psychics*: *An Essay on the Application of Mathematics to the Moral Sciences* (1881), which was clearly written in a state of considerable enthusiasm for his new subject. This slim volume of 150 pages was known only to a small group of experts. Marshall's review began, 'this book shows clear signs of genius, and is a promise of great things to come' (Whitaker 1975, p. 265). Jevons began by stating that 'whatever else readers of this book may think about it, they would probably all agree that it is a very remarkable one' (1881, p. 581). It was not until the middle of the 20th century that many of its central ideas began to be more fully appreciated.

Part 1 of *Mathematical Psychics* (1881, pp. 1–15) was devoted mainly to a justification of the use of mathematics in economics where precise data are not available. There is probably no other 'apology' in the whole of economic literature which compares with Edgeworth's plea for the application of mathematics. For example, when considering individual utility maximization:

> Atoms of pleasure are not easy to distinguish and discern; more continuous than sand, more discrete than liquid; as it were nuclei of the just-perceivable, embedded in circumambient semi-consciousness. We cannot count the golden sands of life; we cannot number the 'innumerable smile' of seas of love; but we seem to be capable of observing that there is here a greater, there a less, multitude of pleasure-units; mass of happiness; and that is enough. (1881, pp. 8–9)

Great stress was placed on comparison with Lagrange's 'principle of least action' in examining the overall effects produced by the interactions among many particles. The connection with Edgeworth's analysis of competition, involving interaction among a large number of competitors to produce a determinate rate of exchange, is central here. The fact that in the natural sciences so much could be derived from a single principle was important for both Jevons and Edgeworth. But Edgeworth took this to its ultimate limit in arguing that the comparable single principle in social sciences, that of maximum utility, would produce results of comparable value. Referring to Laplace's massive work, *Mécanique Céleste,* he suggested that:

> 'Mécanique Sociale' may one day take her place along with 'Mécanique Celeste' [sic], throned each upon the double-sided height of one maximum principle, the supreme pinnacle of moral as of physical science … the movements of each soul, whether selfishly isolated or linked sympathetically, may continually be realising the maximum energy of pleasure, the Divine love of the universe. (1881, p. 12)

Jevons's work in the *Theory of Political Economy* involved the application of very basic mathematics and of psychological research to the analysis of exchange in competitive markets. In addition to this direct stimulus, Edgeworth was also influenced by an anonymous review of Jevons's book in the *Saturday Review* (1871).

The crucial development following Edgeworth's contact with Jevons was not simply the realization that mathematics could be used to examine equilibrium in exchange. Rather, it was that in his analysis Jevons explicitly assumed, through his 'law of indifference', that all individuals take the equilibrium prices as given, that is, outside their control. In using this law as 'one of the central pivots of the theory', Jevons stated that, 'there can only be one ratio of exchange of one uniform commodity at any moment' (1871, p. 87). His theory was explicitly limited to the static equilibrium conditions. He deliberately excluded the role of the number of competitors from his analysis via the awkward notion of the 'trading body', following correspondence with Fleeming Jenkin (1833–1885), who raised the question of indeterminacy with just two traders; see Black (1977, iii, pp. 166–78). Jenkin could not see why two isolated individuals should accept the price-taking equilibrium, whereas Jevons wished to consider the behaviour of two typical individuals in a large market.

In a section on 'Failure of the Laws of Exchange', Jevons discussed cases in which some indeterminacy would result. His most notable example was of house sales, where it was suggested that indeterminacy would result from the discrete nature of the good being exchanged. The *Saturday Review* article took exception to this, suggesting that indeterminacy 'is really owing in our opinion to the assumed absence of competition' (see Black 1981, p. 157). The stress

on indeterminacy was also influenced by Marshall's discussion of wage bargaining: Edgeworth (1881, p. 48 n.1) referred to Thornton's comparison of the determination of prices in Dutch and English auctions, and cited Alfred and Mary Paley Marshall's joint book on the *Economics of Industry* (1879).

It was this gap in Jevons's analysis that Edgeworth set out to fill. His achievement was to show the conditions under which competition between buyers and sellers, through a barter process, leads to a 'final settlement' which is equivalent to one in which all individuals act independently as price takers. As he later stated (1925, p. 453), 'the existence of a uniform rate of exchange between any two commodities is perhaps not so much axiomatic as deducible from the process of competition in a perfect market'.

Exchange and Contract

Having argued that 'the conception of Man as a pleasure machine may justify and facilitate the employment of mechanical terms and Mathematical reasoning in social science' (1881, p. 15). Edgeworth moved on to the analysis of the 'economical calculus', the starting point of which was the assumption that 'every agent is actuated only by self-interest' (1881, p. 16).

In modern economic analysis the analytical tools invented by Edgeworth in 1881, such as the indifference map and the contract curve, are now used in a vast range of contexts. They were introduced by Edgeworth to examine the nature of barter among individuals. He wanted to see if a determinate rate of exchange would be likely to result in barter situations where it is assumed only that individuals wish to maximize their own utility, considered solely as a function of their own consumption. With full knowledge of individuals' utility functions, and their initial endowments of goods, would it be possible to work out a 'determinate' rate of exchange at which trade would take place? Edgeworth's direct statement of the problem is as follows:

> The PROBLEM to which attention is specially directed in this introductory summary is: How far contract is indeterminate – an inquiry of more than theoretical importance, if it show not only that indeterminateness tends to [be present] widely, but also in what direction an escape from its evils is to be sought. (1881, p. 20)

Edgeworth began his analysis of this problem by taking the simplest case of two individuals exchanging fixed quantities of two goods. The basic framework is that described by Jevons, where the first individual holds all of the initial stocks of the first good, and the second individual holds all the stocks of the second good. He wrote the utility functions of each individual in terms of the amounts exchanged rather than consumed, using the general utility function ('utility is regarded as a function of the two variables, not the sum of two functions of each', 1881, p. 104). He then immediately defined the contract curve and indifference curves, in that order.

In the sentence which follows Edgeworth's introduction of the general utility function, he raised the question of the equilibrium which may be reached with 'one or both refusing to move further'. In barter the conditions of exchange must be reached by voluntary agreement, or contract, between the two parties, and of course it is fundamental that no egoist would agree to a contract which would make him worse off than before the exchange. The question thus concerns the nature of the settlement reached by two contracting parties. He immediately answered that contract supplies only part of the answer so that 'supplementary conditions . . . supplied by competition or ethical motives' are required, and then wrote the equation of his famous contract curve (1881, pp. 20–1).

The problem of obtaining the equilibrium values of x and y which, 'cannot be varied without the consent of the parties to it' was stated as follows: 'It is required to find a point (x, y) such that, in whatever direction we take an infinitely small step, $[U_A]$ and $[U_B]$ do not increase together, but that, while one increases, the other decreases' (1881, p. 21). The locus of such points 'it is here proposed to call the contract-curve'. Edgeworth's alternative derivations of the contract curve

involved the movement, from an arbitrary position, along one person's indifference curve; 'motion is possible so long as, one party not losing, the other gains' (1881, p. 23). He thus used the Lagrange multiplier method of maximizing one person's utility subject to the condition that the other person's utility remains constant.

In the diagram drawn by Edgeworth (1881, p. 28) he did not use a box construction. Furthermore the only indifference curves shown fully were those which each individual is able to reach in isolation, and which therefore specify the limits beyond which each is not prepared to move. Also part of the offer or reciprocal demand curves of each individual were drawn on the same diagram, although they were not defined until ten pages later.

After presenting the results for the two-person two-good case, Edgeworth (1881, p. 26) examined the contract curve in the case where three individuals exchange three goods, stated that it is given by the 'eliminant', and then gave three lines of three sets of partial derivatives. In fact, the contract curve in this context is defined by $\left|\frac{\partial U_i}{\partial x_j}\right|$, where $\frac{\partial U_i}{\partial x_j}$, is the marginal utility of person i with respect to good j, but Edgeworth did not use the modern notation for determinants and did not set the Jacobian equal to zero. This early use of determinants in economics would probably have confused many of his readers.

The Problem of Indeterminacy

The concepts of indifference curves and the contract curve therefore help to specify a range of 'efficient exchanges' of goods between individuals. The essential feature of the analysis from Edgeworth's point of view is precisely that there is a range rather than a unique point: 'the settlements are represented by an indefinite number of points' (1881, p. 29). At any particular settlement, the rate of exchange is expressed simply in terms of the amount of one good which is given up in order to obtain a specified amount of the other good. Hence the existence of a range of efficient contracts means that the rate of exchange is 'indeterminate'. The rate of exchange achieved in practice will thus depend to a large extent on bargaining strength. It was this result which led Edgeworth to make his often quoted remark that 'an accessory evil of indeterminate contract is the tendency, greater than in a full market, towards dissimulation and objectionable arts of higgling' (1881, p. 30).

Edgeworth argued that his analysis of indeterminacy in contract between two traders could be applied to a very wide variety of contexts. In particular, the tendency of large groups to form 'combinations', as in the case of trade unions and employers' associations, would serve to increase the extent of indeterminacy. The general applicability of his analysis of contract and indeterminacy was summarized by Edgeworth as follows:

> What it has been sought to bring clearly into view is the essential identity (in the midst of diversity of fields and articles) of contract; a sort of unification likely to be distasteful to those excellent persons who are always dividing the One into the Many, but do not appear very ready to subsume the Many under the One. (1881, p. 146; Plato's expression 'the one in the many' was later used by Marshall as the motto for his 1919 book on *Industry and Trade*.)

Having shown the possibilities of indeterminacy, Edgeworth then went on to show how 'the escape from its evils' requires either competition or arbitration.

Competition and the Number of Traders

The central question which Edgeworth was trying to resolve in the second part of *Mathematical Psychics* was that of the conditions necessary to remove the indeterminacy which exists in the case of barter between two traders. The question naturally arises as to the extent to which this indeterminacy is the result of the absence of competition in the simple two-person market. Edgeworth thus quickly moved on to the introduction of further traders.

In Edgeworth's earlier problem of two traders exchanging two goods, the definition of a range of

efficient exchanges (along the contract curve) is of course analytically separate from the question of whether or not two isolated traders would actually reach a settlement on the contract curve. However, these two aspects were not clearly separated by Edgeworth because at the beginning of his analysis he introduced his stylized description of the process of barter: this is the famous 'recontracting' process. Edgeworth did not wish to assume that individuals initially have perfect knowledge. Instead, he supposed that, '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, but sufficiently approximate to existence or tendency for the purposes of abstract science' (1881, p. 18). The knowledge of the other traders' dispositions and resources could be obtained by the formation of tentative contracts which are not assumed to involve actual transfers, and can be broken when further information is obtained. Edgeworth introduced this in typical style:

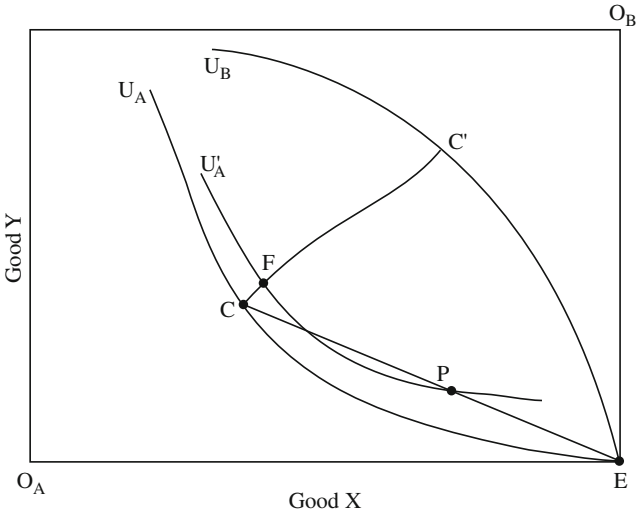
> 'Is it peace or war?' asks the lover of 'Maud', of economic competition, and answers hastily: it is both, pax or pact between contractors during contract, war, when some of the contractors without the consent of others recontract. (1881, p. 17; the allusion here is to Alfred Tennyson's poem *Maud*: *A Monodrama,* part 1, verse VII.)

An important role of the recontracting process is thus to disseminate information among traders. It allows individuals who initially agree to a contract, which is not on the contract curve, to discover that an opportunity exists for making an improved contract according to which at least one person gains without another suffering.

However, the real importance of the recontracting process lies in the fact that it allows for Edgeworth's analysis of the role of the number of individuals in a market. With numerous individuals, the recontracting process makes it possible to analyse the use of collusion among some of the traders. Individuals are allowed to form coalitions in order to improve bargaining strength. Recontracting enables the coalitions to be broken up by outsiders who may attract members of a group away with more favourable terms of exchange.

Edgeworth's analysis was extremely terse and the following discussion does not therefore follow his own presentation. The analysis begins by introducing a second person A and a second person B. The new traders are assumed to be exact replicas of the initial pair, with the same tastes and endowments. This simplification is useful because the dimensions of the Edgeworth box and the utility curves are identical for each pair of traders. Hence, it enables the same diagram to be used as in the case when only two traders are considered in isolation. Two basic points can be stated immediately. First, in the final settlement all individuals will be at a common point in the Edgeworth box. Second, the settlement must be on the contract curve. The first point arises because if two individuals have identical tastes then their total utility is maximized by sharing their resources equally. It is useful to consider other types of contract which will eventually be broken, in order to illustrate the way in which the introduction of additional traders provides a role for some kind of competitive process.

The major question at issue is whether the range of indeterminacy along the contract curve is reduced by the addition of these traders. Consider Fig. 1 and suppose that when A_1 and B_1 are trading independently of A_2 and B_2, trader B_1 has all the bargaining power and is able to appropriate all the gains from trade by pushing A_1 to the limit of the contract curve at point C. Suppose also that the same applies to A_2 and B_2. If the two pairs of traders are then able to communicate with each other, A_2 can now simply refuse to trade with B_2 at C. With no transaction costs, A_2 was previously indifferent between trading at C and consuming at the endowment point, E. This endowment position is effectively the 'threat point' of the As: it is the position in which they would find themselves if the bargaining process were to break down. But A_2 no longer needs to remain in isolation after refusing to trade with B_2, and instead can trade with A_1, after A_1 has traded with B_1 at C and has therefore obtained some of good Y. The two As can share their stocks of X and Y equally, arriving at point P; such an equal division maximizes their total utility.



Edgeworth, Francis Ysidro (1845–1926), Fig. 1 Two pairs of traders

By reaching point *P,* halfway between *C* and *E,* the convexity of the indifference curves implies that they are both better off than anywhere on the no-trade indifference curve. The two As would be on a higher common indifference curve, and thus better-off, if they could consume at a point along the *CPE* which is to the north-west of point *P* However, they do not have enough resources to move beyond the halfway point *P.*

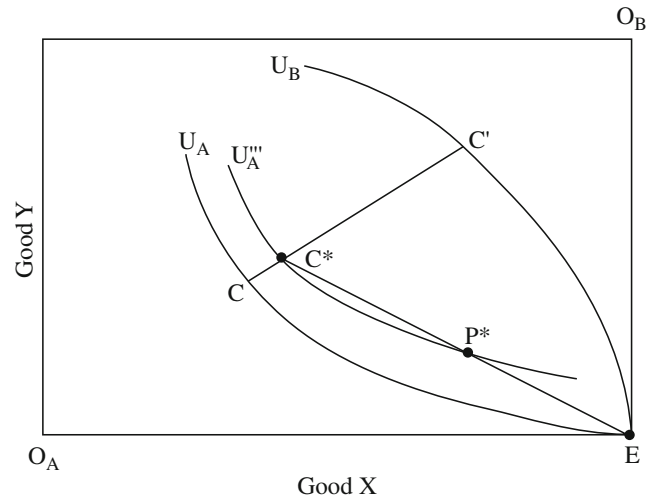
Trader B_2, who has been isolated, cannot prevent such a bargain. Thus B_1 is at *C*, both *A*s are at *P* and B_2 is at the initial endowment point *E*. In this situation B_1 has no incentive to change, but B_2 has a strong incentive to offer a better deal to one of the *A*s than the one offered by trader B_1. So long as B_2 offers one of the *A*s, say A_2, a trade on the contract curve which allows A_2 to reach a higher indifference curve than U'_A, the initial agreement with B_1 will be broken and recontracting will take place.

The implication is that the ability of the *A*s to turn to someone else, rather than deal with a single trader, means that the *B*s now compete against each other. However, trader B_1, who cannot prevent the recontracting, has an incentive to make yet a better offer. Hence, the recontracting process continues. The stylized process of recontracting with the two *B*s competing against each other will produce a final settlement at the point *C** in Fig. 2. This has the property that the indifference curve U'''_A passes through *C** and *P**, where *P** is halfway between *C** and *E*. This means that the two *A*s are indifferent between *C** and *P**, and since they cannot both reach any point between *C** and *P** along the line *C** *E,* they are unable to improve on *C**. Hence there is no need to leave one of the *B*s in isolation and the two *B*s will trade with the two *A*s at point *C**.

This argument has shown that at the final settlement all traders are at a common point on the contract curve and the limit has moved inwards along the old contract curve. The analysis can be repeated by starting with an alternative situation whereby the *A*s are initially assumed to be able to appropriate all the gains from trade. The point *C*′ would then no longer qualify as a point on the new contract curve. The introduction of the additional pair of traders means that the contract curve shrinks, and the range of indeterminacy involved in barter is correspondingly reduced.

The extent to which the contract curve shrinks when the additional pair of traders is introduced is influenced by the fact that the *A*s cannot get further than halfway along a ray from a point on the contract curve to the endowment position.

However, if there are three pairs of *As* and three pairs of *B*s, the repetition of the above analysis involves two of the *A*s dealing with two of the *B*s at a point on the contract curve. The two *A*s then share their resources equally with the remaining



Edgeworth, Francis Ysidro (1845–1926), Fig. 2 The new limit to the contract curve

A while the third *B* is isolated. The *A*s are able to consume together at a point which is two-thirds of the way along the ray from the initial endowment position to the point on the contract curve where the trade involving the two *A*s and two *B*s takes place.

With *N* pairs, the *A*s can reach a proportion $\frac{(N-1)}{N}$ of the way from the endowment point to the contract curve. Thus as *N* increases, the values of *k* approaches unity. This means that the *A*s can reach all the way from *E* to the contract curve, so that the final settlement must be such that the indifference curve is tangential to the ray from the origin. *A* final settlement with many traders is therefore shown in Fig. 3 as point *P* on the contract curve. The effect of working in from the point C' would lead to an equivalent result for an indifference curve of the *B*s, shown as U_B^*.

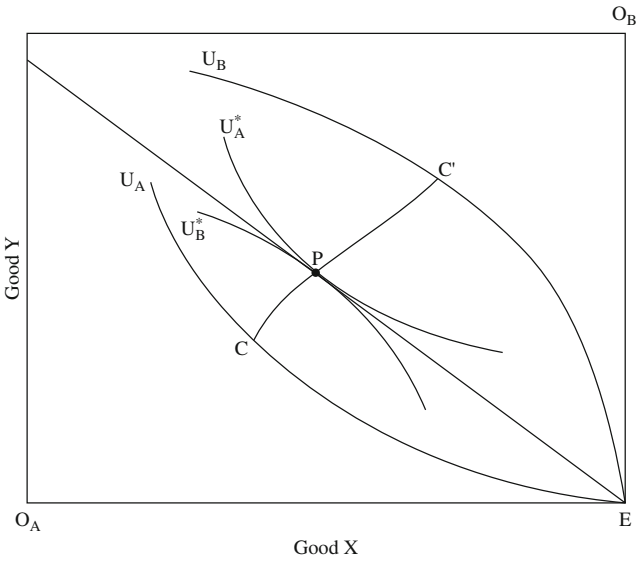
The result is that the final settlement looks just like a price-taking equilibrium. The figure illustrates the case where there is a single price-taking equilibrium. If there are multiple equilibria, the recontracting process causes the number of final settlements, with sufficiently large *N,* to shrink to the number of price-taking equilibria. (For discussion of utility functions involving multiple equilibria, and comparison of bargaining, competitive and utilitarian solutions, see Creedy 1994a.) This argument relating to the shrinking contract curve, first established by Edgeworth, is often referred to as the *limit theorem.*

After Edgeworth's terse discussion, he stated:

> If this reasoning does not seem satisfactory, it would be possible to give a more formal proof; bringing out the important result that the common tangent to both indifference curves . . . is the vector from the origin. (1881, p. 38)

The price-taking solution is necessarily on the contract curve. This gives rise to what is now referred to as the 'first fundamental theorem' of welfare economics -that a price-taking equilibrium is Pareto efficient. Furthermore, the use of price-taking provides a considerable reduction in the amount of information required by traders when compared with the recontracting process. Given an equilibrium set, individuals need to know only the prices of goods, whereas in the recontracting process they have to learn a considerable amount of information about other individuals' preferences and endowments. But Edgeworth placed more stress on the equivalence of the competitive price-taking solution with a recontracting barter process involving large numbers.

Given that coalitions among traders are allowed in the recontracting process, a price-taking equilibrium cannot be blocked by a coalition of traders. In this sense the competitive equilibrium is robust. The argument that a complex process of bargaining among a large number of individuals produces a result which replicates a price-taking equilibrium, allowing for the free flow of



Edgeworth, Francis Ysidro (1845–1926), Fig. 3 Final settlement with many traders

information using recontracting and enabling coalitions of traders to form and break up, is an important result that is far from intuitively obvious. The recontracting process can be said to represent a competitive process, and the contract curve shrinks essentially because of the competition between suppliers of the same good, although it is carried out in a barter framework in which explicit prices are not used (although rates of exchange are equivalent to price ratios).

The price-taking equilibrium, in contrast, does not actually involve a competitive process. Individuals simply believe that they must take market prices as given and outside their control. They respond to those prices without any reference to other individuals. But the result is that the price-taking equilibrium looks just like a situation in which all activity is perfectly coordinated.

Edgeworth suggested that similar results apply when some of the assumptions are relaxed. Thus, 'when we suppose plurality of natures as well as persons, we have to suppose a plurality of contract-curves . . . Then, by considerations analogous to those already employed, it may appear that the quantity of final settlements is diminished as the number of competitors is increased' (1881, p. 40). He then briefly considered different numbers of *A*s and *B*s, concluding that 'the theorem admits of being extended to the general case of unequal numbers and natures' (1881, p. 43). However, some of the results do not hold in the general case; for example, equality within the group of *A*s no longer holds when there are unequal numbers of *A*s and *B*s. A considerable number of articles have been written, since the late 1950s, examining various aspects of the Edgeworth recontract model under different assumptions.

Reciprocal Demand Curves

It has been mentioned that Edgeworth included in his diagram (1881, p. 28) the reciprocal demand curve, or offer curve, of each individual, although such curves were then called 'demand-and-supply curves'. Edgeworth mentioned them only briefly in the text (1881, p. 39), but the lack of emphasis is understandable since in imperfect competition they are not relevant. Edgeworth's contribution was to provide the basic 'analytics' of the offer curve in terms of indifference curves, whereby it is 'the locus of the point where lines from the origin touch curves of indifference' (1881, p. 113).

When there is a lack of competition, giving rise to indeterminacy, there is nothing to ensure that individuals will trade on their offer curves and, as Edgeworth argued, 'the conceptions of demand and supply at a price are no longer appropriate'

(1881, p. 31). It is this general preference, in favour of the analysis of barter in noncompetitive situations, to which Marshall objected and which led to the controversy discussed below.

The Utilitarian Calculus

Having shown how indeterminacy can be removed by increasing the number of traders, Edgeworth turned to consider the role of arbitration in resolving the conflict between traders, in a 'world weary of strife' (1881, p. 51). The principle of arbitration examined was, not surprisingly, the utilitarian principle, which Edgeworth had earlier used to examine the optimal distribution. However, the new context of indeterminacy led him to a deeper justification of utilitarianism as a principle of distributive justice. Having arrived at this new link between 'impure' (egoistic) and 'pure' utilitarianism, Edgeworth had only to reorientate his earlier analysis of optimal distribution, contained in his paper in *Mind* of 1879.

The need for arbitration with indeterminacy had been stated by Jevons as follows:

> The dispositions and force of character of the parties ... will influence the decision. These are motives more or less extraneous to a theory of economics, and yet they appear necessary considerations in this problem. It may be that indeterminate bargains of this kind are best arranged by an arbitrator or third party. (1871, pp. 124–5)

Edgeworth's statement of the same point was as usual rather less prosaic: 'The whole creation groans and yearns, desiderating a principle of arbitration, and end of strifes' (1881, p. 51). Edgeworth argument involved two steps. First, he showed that the principle of utility maximization places individuals on the contract curve, because the first-order conditions are equivalent to the tangency of indifference curves.

> It is a circumstance of momentous interest that one of the in general indefinitely numerous settlements between contractors is the utilitarian arrangement ... the contract tending to the greatest possible total utility of the contractors. (1881, p. 53)

Edgeworth recognized that this result was not sufficient to justify the use of utilitarianism as a principle of arbitration. It is only a necessary condition of a principle of arbitration that it should place the parties somewhere on the contract curve. Edgeworth's justification for utilitarianism as a principle of justice, comparing points along the contract curve, was as follows:

> Now these positions lie in a reverse order of desirability for each party; and it may seem to each that as he cannot have his own way, in the absence of any definite principle of selection, he has about as good a chance of one of the arrangements as another... both parties may agree to commute their chance of any of the arrangements for ... the utilitarian arrangement. (1881, p. 55)

The important point to stress about this statement is that Edgeworth clearly viewed distributive justice in terms of choice under uncertainty. He argued that the contractors, faced with uncertainty about their prospects, would choose to accept an arrangement along utilitarian lines. A crucial component of this argument, also clearly stated by Edgeworth in this quotation, is the use of equal a priori probabilities.

The importance to him of this new justification of utilitarianism cannot be exaggerated. Indeed the whole of *Mathematical Psychics* seems to be imbued with a feeling of excitement generated by his discovery of a justification based on a 'social contract'. This provided the crucial link between 'impure' and 'pure' utilitarianism in a more satisfactory way than his earlier appeal to evolutionary forces.

Edgeworth believed that he had provided an answer to an age-old question, stating 'by what mechanism the force of self-love can be applied so as to support the structure of utilitarian politics, neither Helvetius, nor Bentham, nor any deductive egoist has made clear' (1881, p. 128). Nevertheless this argument was neglected until restatements along similar lines were made by Harsanyi (1953, 1955) and Vickrey (1960). The maximization of expected utility, with each individual taking the a priori view that any outcome is equally likely, was shown to lead to the use of a social welfare function which maximizes the sum of individual utilities. This approach is now usually described as 'contractarian neo-utilitarianism'.

In discussing the utilitarian solution as a principle of arbitration in indeterminate contract, Edgeworth did not clearly indicate in 1881 that the utilitarian solution of maximum total utility could specify a position which makes one of the parties worse off than in the no-trade situation. This was nevertheless later made explicit when, after proposing arbitration along utilitarian lines, he added 'subject to the condition that neither should lose by the contract' (1925, ii, p. 102). This possibility of course depends largely on the initial endowments of the individuals.

Later Work in Economics

After the publication of *Mathematical Psychics*, Edgeworth concentrated increasingly on mathematical statistics, in particular on the problem of statistical inference, but, following his appointment to the Drummond Chair at Oxford, Edgeworth again made important contributions to economics, although this work mainly involved reactions to, and discussions arising from, the later work of other authors.

Demand and Exchange

In the *Principles of Economics* (1890, Appendix F) Marshall included a brief discussion of Edgeworth's analysis of barter, and produced a figure showing the contract curve. During the following year, in the course of a review written in Italian (translated in Edgeworth, 1925, ii, pp. 315–19), Edgeworth criticized Marshall for not having dealt sufficiently with the problem of indeterminacy. The basic problem was that Marshall, using a model in which a series of trades are allowed to take place at disequilibrium prices, believed he had shown that prices will eventually settle at the price-taking equilibrium. However, the argument was not transparent.

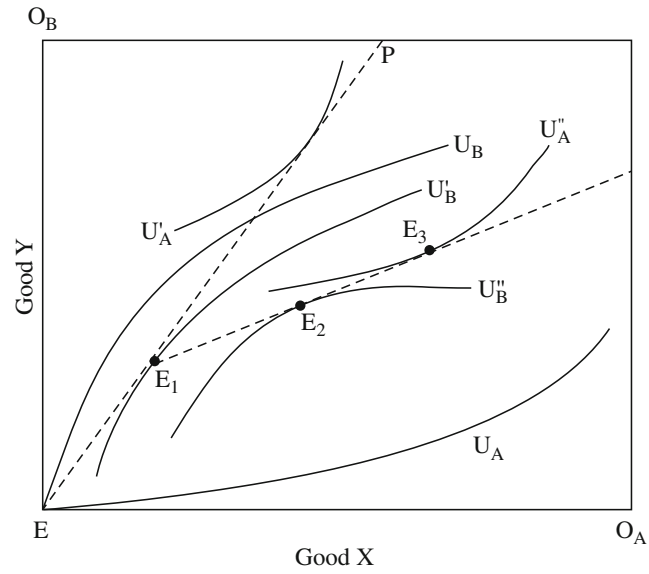
The adjustment process involves moving from the initial endowment point in a series of trades, where trading at 'false' prices is allowed at each step. The process must conclude with both individuals at a point on the contract curve. A feature of the process is the assumption that each stage or iteration of the sequence involves Pareto improvements: individuals trade only if it makes them better off. Furthermore, it involves trading at the 'short end' of the market, that is, the minimum of supply and demand. This arises from the impossibility of forcing any individual either to buy or sell more than desired at any price.

An example of two disequilibrium trades is shown in Fig. 4, where the endowment moves from E to E_1, and then to E_2. With a price line represented by EP, there is an excess supply of good X as person A tries to reach the indifference curve U'_A and person B wishes to reach U'_B. Trade takes place at E_1, the short end of the market. Point E_1 then becomes the new endowment point. At the second trading stage, the price of X must be lowered to induce person B to purchase more. At a price represented by the line E_1P_1 through the new endowment point, the excess supply is lower than formerly and trade takes place at E_2. Comparing U'_A and U'_B with U''_A and U''_B with UA and UB respectively, it can be seen that E_2 is a Pareto improvement relative to E_1. It is also clear that person A is better off the slower the fall is in the price of X relative to Y at each stage.

The combination of Pareto-efficient moves at each stage and an adjustment process such that an excess supply leads to a price reduction, and vice versa, produces a stable process that converges to an equilibrium somewhere on the contract curve. (This type of sequence of disequilibrium trades was later used by Launhardt; see Creedy 1994b.)

The basic problem was that Marshall believed that his assumption of an additive utility function, combined with the assumption that the marginal utility of one good is constant for both individuals, guaranteed a determinate price, if the good having constant marginal utility was money. Indeed, this case was mentioned by Edgeworth (see 1925, ii, p. 317 n.1). The contract curve is a straight line parallel to the y axis (where this good is the one with constant marginal utility), along which the rate of exchange is constant. So the equilibrium price does not depend on the sequence of trades. However, Edgeworth's point was that the total amount spent on good x remains indeterminate.

There was a later, though much milder, disagreement between Marshall and Edgeworth



Edgeworth, Francis Ysidro (1845–1926), Fig. 4 Disequilibrium trades

over the so-called Giffen good. In a book review, Edgeworth argued that, 'even the milder statement that the elasticity of demand for wheat may be positive, though I know it is countenanced by high authority, appears to me so contrary to a priori probability as to require very strong evidence (1909, p. 104). The 'authority' was of course Marshall (1890, p. 132), who replied directly to Edgeworth that, 'I don't want to argue ... But... the matter has not been taken quite at random' (Pigou 1925, p. 438). Marshall gave a numerical example involving a journey travelled by two methods, where the distance travelled by the cheaper and slower method must increase when its price increases. For further details, see Creedy (1990).

It has been mentioned that Edgeworth introduced the generalized utility function. An implication is that it allows for complementarity, although Edgeworth did not explicitly consider this in 1881. The first formal definition of complementarity is attributed to Auspitz and Lieben, and it was used by Edgeworth in his paper on the pure theory of monopoly, and also by Pareto: this amounts to what is now called 'gross' complementarity, defined in terms of cross-price elasticities. It is also sometimes referred to, using the initials of the four people mentioned above, as ALEP complementarity.

The first major criticism came from Johnson (1913), who pointed out that the criterion was not invariant with respect to monotonic transformations of the utility function. His treatment was extended by Hicks and Allen (1934), so that the modern definition involves 'net' complements in terms of compensated price changes. There is no symmetry between gross substitutes and complements as only the matrix of (compensated) substitution elasticities is assumed to be symmetric.

Monopoly and Oligopoly

In a paper first published in Italian in 1897, and not translated until the collected *Papers* (1925), Edgeworth examined several problems relating to monopoly. He began his discussion with Cournot's (1838) example of the 'source minérale' in which there are 'two monopolists' (that is, duopolists), each owning a spring of mineral water. It would be natural for Edgeworth to expect an indeterminate price in this 'small numbers' context. Cournot had arrived at a determinate solution for price and output, but Edgeworth showed that 'when two or more monopolists are dealing with competitive groups, economic equilibrium is indeterminate' (1925, p. 116). The daily output from each spring was assumed to be limited to identical fixed amounts, delivery costs were zero and all consumers had the same demand

curve (purchasing one unit only of output). Hence demand is $n(1 - p)$ where n is the number of customers and p is the price. Cournot's solution was that the price would be $p = 1/4$, but Edgeworth argued that one of the 'monopolists' had an incentive to raise the price back to $p = 1/2$, which is the revenue maximizing price, so that there is not a determinate price. He argued that:

> at every stage ... it is competent to each monopolist to deliberate whether it will pay him better to lower his price against his rival as already described, or rather to raise it to a higher . . . for that remainder of customers of which he cannot be deprived by his rival. . . . Long before the lowest point has been reached, that alternative will have become more advantageous than the course first described' (1925, p. 120)

Edgeworth went on to say 'the matter may be put in a clearer light', and he then defined what are now called the reaction curve and isoprofit lines (in that order) for variations in prices. However, it was not until Bowley's (1924) discussion that these matters began to be presented in a more transparent manner.

Edgeworth then considered the case of complementary demand within the context of 'bilateral monopoly', where the two goods are demanded in fixed proportions for use in the production of a further article. An interesting feature is that he wrote the equations of the reaction curves and explicitly dealt with what are now called conjectural variations, reflecting the extent to which one duopolist is expected to change price in response to changes made by the second duopolist. In discussing this problem Edgeworth also introduced the further important concept of the 'saddle point', which he called the 'hog's back', clearly indicating its importance for stability.

The No-Profit Entrepreneur

Walras (1874, p. 225) had introduced the concept of the entrepreneur who neither gains nor loses. This result applied only to the competitive equilibrium, where there are no incentives for entrepreneurs to enter any industry. This does not of course mean that there are no profits, in the accounting sense, since the returns to homogeneous units of inputs of organization and management services are subsumed in the costs of the firm.

Edgeworth's criticisms of this concept of the no-profit entrepreneur, reproduced in his *Papers* (1925), recognized that with Walras's assumptions there was nothing illogical about the argument. The theory simply means that nothing remains 'after the entrepreneur has paid a normal salary to himself' (1925, pp. 26, 30). Furthermore, 'if [the general expenses] are taken into account, the argument becomes a fortiori. For why should not a substantial remuneration for the entrepreneur be included in the general expenses of the business' (1925, ii, p. 469). Edgeworth's difference with Walras was to some extent 'only verbal', but he was also unhappy with the idea that entrepreneurship is homogeneous and divisible.

The Theory of Taxation

In the 1890s Edgeworth produced two surveys of considerable importance. These surveys, of the pure theory of taxation and of the pure theory of international values, were both published in the *Economic Journal* and subsequently reproduced (with alterations) in his *Papers* (1925, vol. ii). Each survey consisted of three separate parts, and displayed a staggering breadth of knowledge and command of the subject. They represent his most serious attempts to produce any kind of synthesis of a branch of economic literature. Edgeworth began his survey with the rather strong statement that 'the science of taxation comprises two subjects to which the character of pure theory may be ascribed; the laws of incidence, and the principle of equal sacrifice' (1925, p.64). He then considered a variety of special cases and contexts of tax incidence. The basic framework for incidence analysis was the simple partial equilibrium approach, still used in many basic textbooks, in which the incidence depends on the relative values of supply and demand elasticities.

The basic approach to incidence analysis actually stemmed from the important paper by Jenkin (1871). It suggests that in general the price of the taxed good will either remain constant (in the extreme case of inelastic supply) or will increase. However, this result ignores interrelationships

among commodities. Edgeworth showed that, when such interrelationships are explicitly allowed, there are some circumstances in which the price of the taxed good will actually fall. When discussing this 'paradox', Edgeworth reproduced his argument which had in fact been explored in more detail in his paper on monopoly, published in Italian in the same year (translated in Edgeworth, 1925, *i,* pp. 111–42). Edgeworth first stated his 'tax paradox' in the following terms:

> when the supply of two or more correlated commodities – such as the carriage of passengers by rail first class or third class – is in the hands of a single monopolist, a tax on one of the articles – e.g. a percentage of first class fares – may prove advantageous to the consumers as a whole. . . . The fares for all the classes might be reduced. (1925, p. 139)

Edgeworth regarded this result as an example of a situation where, 'the abstract reasoning serves as a corrective to what has been called the "metaphysical incumbus" of dogmatic laisser faire' (1925, *i,* p. 139; see also 1925, ii, pp. 93–4). Essentially the two commodities must be substitutes in consumption and production, and the result is partly brought about by the fact that the monopolist has an incentive to increase the supply of the untaxed commodity. Edgeworth also recognized that the result could occur in competitive markets (see 1925, p. 63). As with many of Edgeworth's original results, this tax paradox was not a subject of continuous development. Its main practical importance perhaps arises from the fact that in the early 1930s it attracted the attention of Hotelling (1932). For further discussion of the paradox, see Creedy (1988).

The section of the taxation survey which attracted most immediate attention was Edgeworth's discussion of the various 'sacrifice' theories of the distribution of the tax burden, and his qualified support for progressive taxation. Edgeworth's attitude to taxation was similar to that of the major classical economists in that he rejected a benefit approach, on the argument that taxation is not an economic bargain governed by competition. Thus in his view the problem was to determine 'the distribution of those taxes which are applied to common purposes, the benefits whereof cannot be allocated to particular classes of citizens' (1925, p. 103). A principle of justice is thus required. His approach can be seen as marking a crucial stage in the transition towards a 'welfare economics' view of public finance, rather than using a special set of 'tax maxims' such as the famous criteria laid down by Adam Smith.

Not surprisingly, Edgeworth (1925, p. 102) argued along neo-contractarian lines set down in *Mathematical Psychics* that the utilitarian arrangement would be accepted by individuals uncertain of their own prospects and taking an equal a priori view of the probabilities. He suggested that

> each party may reflect that, in the long run of various cases ... of all the principles of distribution which would afford him now a greater, now a smaller proportion of the sum-total utility obtainable . . . the principle that the collective utility should be on each occasion a maximum is most likely to afford the greatest utility in the long run to him individually

Having established the use of utilitarianism as a principle of distributive justice, Edgeworth then succinctly stated the main argument:

> The condition that the total net utility procured by taxation should be a maximum then reduces to the condition that the total disutility should be a minimum . . . it follows in general that the marginal disutility incurred by each taxpayer should be the same. (1925, p. 103)

The implication is that, if all individuals have the same cardinal utility function, after-tax incomes would be equalized. Edgeworth also clearly recognized that, if there is considerable dispersion of pre-tax incomes relative to the total amount of tax to be raised, where there is 'not enough tax to go around' (1925, ii, p. 103), the equimarginal condition cannot be fully satisfied unless there is a 'negative income tax' which raises the incomes of the poorest individuals to a common level. Thus, 'the acme of socialism is for a moment sighted' (1925, p. 104). But Edgeworth immediately considered the practical limitations to such high progressive taxation. The following quotation illustrates one of Edgeworth's favourite metaphors, his respect for Sidgwick, his attitude to authority, his views on utilitarianism and the

applicability of pure theory, and of course his unmistakable style:

> In this misty and precipitous region let us take Professor Sidgwick as our chief guide. He best has contemplated the crowning height of the utilitarian first principle, from which the steps of a sublime deduction lead to the high tableland of equality; but he also discerns the enormous interposing chasms which deter practical wisdom from moving directly towards that ideal. (1925, p. 104)

Among the various limitations, Edgeworth noted differences in individual utility functions, population effects, the disincentives to work, growth of culture and knowledge, savings, and of course the problem of evasion.

International Trade

Edgeworth's survey of the pure theory of international values was in some ways responsible for a change of emphasis in the approach to trade theory, despite the fact that it contained few original analytical contributions. Indeed, he said that, 'Mill's exposition of the general theory is still unsurpassed' (1925, p. 20), and acknowledged further that, 'what is written ... after a perusal of [Marshall's] privately circulated chapters . . . can make no claim to originality' (1925, p. 46). Edgeworth saw trade theory as an application of the general theory of exchange:

> The fundamental principle of international trade is that general theory ... the Theory of Exchange ... which ... constitutes the 'kernel' of most of the chief problems in economics. It is a corollary of the general theory that all the parties to a bargain look to gain by it . . . This is the generalised statement of the theory of comparative cost. (1925, p. 6)

Thus the gains from trade are analogous to the gains from exchange in simple barter and 'It is useful . . . to contemplate the theory of distribution as analogous to that of international trade proper' (1925, p. 19). Hence trade theory is to Edgeworth simply one more application of the general method of *Mathematical Psychics.* In directly applying the theory of exchange to that of trade, Edgeworth was quite content to use community indifference curves without clearly specifying how aggregation might be carried out. He said only that 'by combining properly the utility curves for all the individuals, we obtain what may be called a collective utility curve' (1925, p. 293).

One of Edgeworth's criticisms of Mill (1848) was that the latter took as his measure of the gain from trade the change in the ratio of exchange of exports against imports. Thus Mill in this case 'confounds "final" with integral utility' (1925, p. 22). The same point had in fact been made by Jevons (1871, pp. 154–6). However, Edgeworth, while preferring total utility, admitted that Mill was not otherwise led to serious error in using his own measure.

Edgeworth's survey was, as always, extremely wide-ranging, though for later developments the most interesting parts are concerned with his elucidation of Mill's 'recognition of the case in which an impediment may be beneficial – or an improvement prejudicial – to one of the countries' (1925, p. 9). These cases would now be discussed under the headings of the 'optimal tariff' and 'immiserizing growth'. In the case of an optimal tariff, a country acts as monopolist and imposes a price which enables that country to attain its highest indifference curve, subject to the other country's offer curve. However, this position is not on the contract curve. The detailed specification of the optimum tariff in terms of elasticities had to wait until Bickerdike (1906) and Pigou (1908) and the later revivals of interest in the 1940s. Edgeworth's judgement of Bickerdike was that he had 'accomplished a wonderful feat. He has said something new about protection' (1925, ii, p. 344).

Edgeworth could not of course be expected to support the use of such tariffs in practice. He acknowledged the possibility of retaliation, but also:

> For one nation to benefit itself at the expense of... others is contrary to the highest morality . . . But in an abstract study upon the motion of projectiles in vacuo, *I* do not think it necessary to enlarge upon the horrors of war. (1925, p. 17 n. 5)

The 'highest morality' was, of course, the principal of utilitarianism.

Conclusions

It has been seen that Edgeworth did not begin working and writing in economics until his

mid-30s, but in common with the majority of neoclassical economists he soon pursued an academic career as a professor of economics. Indeed, in a period which saw the rapid and widespread professionalization of the subject Edgeworth held an academic position in England that was regarded as second only to that of Alfred Marshall. In spite of his wide range of reading and sympathies, Edgeworth's work was characterized by the fact that it was virtually all addressed to his fellow professional economists. So uncompromising was he in his view that economics is a very difficult subject offering only remote and nearly always negative policy advice that it may fairly be said that his work was addressed to just a small number of 'fellow travellers' in the rarefied atmosphere of the 'higher regions' of pure theory. However, Edgeworth imposed no geographical limitations, and with his considerable linguistic skills and international sympathies was in contact with the majority of leading economists around the world.

The distinguishing feature of the neoclassical 'revolution' was its emphasis on exchange as the central economic problem. The success of this shift of focus from production and distribution to exchange was closely associated with the fact that it had as its foundation a model based on utility maximization. This allowed for a deeper treatment of the gains from exchange and the wider considerations of economic welfare. Schumpeter summarized the point by stating that utility analysis must be understood in terms of exchange as the central 'pivot' and 'the whole of the organism of pure economics thus finds itself unified in the light of a single principle' (1954, p. 913). This is indeed the context in which Edgeworth's work in economics must be seen. Schumpeter's remark is merely a more prosaic expression of Edgeworth's view quoted above that '"Mechanique Sociale" may one day take her place along with "Mechanique Celeste" [sic], throned each upon the double-sided height of one maximum principle'. The central theme of Edgeworth's work is also clear in his revealing statement, taken from his presidential address to Section *F* of the Royal Society, that:

> It may be said that in pure economics there is only one fundamental theorem, but that is a very difficult one: the theory of bargain in a wide sense. (1925, ii, p. 288)

This perspective helps the major thread which runs through all Edgeworth's work in economics to be seen. His earlier mathematical analysis of the implications of utilitarianism for the optimal distribution, written before he turned to economics, was not only highly original (and esoteric) but laid the foundation for his work in economics. Thus, the transition from *New and Old Methods of Ethics* to *Mathematical Psychics* was not a shift in major preoccupations but rather a change of emphasis. Distribution was then seen as an important concomitant of exchange, so that the analysis of contract became central for Edgeworth. Edgeworth's emphasis on the indeterminacy (the inability of utility maximization alone to determine the rate of exchange, only a range of efficient exchanges) which results from the existence of a small number of traders led him to his path-breaking analysis of the role of numbers in competition, along with the efficiency properties of competitive equilibria.

The analysis of the utilitarian objective as an arbitration rule led Edgeworth directly to his new 'social contract' argument in explaining the acceptance of utilitarianism as a principle of social justice. It was the realization of this new justification of utilitarianism, using his newly developed analytical tools, which generated the excitement that is clearly evident in his first work in economics. While *Mathematical Psychics* developed the techniques of indifference curves and the contract curve within the 'Edgeworth box' – tools which are now ubiquitous in economic analysis – Edgeworth himself was clearly driven mainly by his ability to link the analysis of private contracts in markets to that of a social contract in which utilitarianism is the 'sovereign principle'. The integration of his analysis of barter, and the effects of the introduction of additional traders into the market, with the demonstration that the utilitarian arrangement prescribes a point on the contract curve of efficient exchanges and is acceptable to risk-averse traders, was to Edgeworth nothing short of 'momentous'.

The results are of course highly abstract. In discussing their ultimate value suggested that:

> Considerations so abstract it would of course be ridiculous to fling upon the flood-tide of practical politics … it is at a height of abstraction in the rarefied atmosphere of speculation that the secret springs of action take their rise, and a direction is imparted to the pure foundation of youthful enthusiasm whose influence will ultimately affect the broad current of events. (1881, p. 128)

The intellectual pleasure derived from being able to draw together so many different subjects of analysis, and strands of his enormous range of learning, is clearly evident. However, it is precisely this wide field of vision, combined with the technical level and idiosyncratic style of writing, which made *Mathematical Psychics* so difficult for his contemporaries, and which continue to make the book seem so strange and yet so rewarding to the modern reader.

Selected Works

1876. Mr. Matthew Arnold on Bishop Butler's doctrine of self love. *Mind* 1: 570–571.

1877. *New and old methods of ethics: Or 'physical ethics' and 'methods of ethics'*. Oxford: Parker.

1879. The hedonical calculus. *Mind* 4: 394–408.

1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. London: Kegan Paul.

1887. *Metretike, or the method of measuring probability and utility*. London: Temple.

1909. Review of free trade in being. *Economic Journal* 19: 104–105.

1925. *Papers relating to political economy*, 3 vols. London: Macmillan, for the Royal Economic Society.

Acknowledgment I am grateful to Denis O'Brien and Steven Durlauf for comments on an earlier draft of this article.

Bibliography

Bickerdike, C. 1906. The theory of incipient taxes. *Economic Journal* 16: 529–535.

Black, R., ed. 1977. *Papers and correspondence of William Stanley Jevons*. London: Macmillan, for the Royal Economic Society.

Black, R., ed. 1981. *Papers and correspondence of William Stanley Jevons*, Papers on political economy. Vol. VII. London: Macmillan, for the Royal Economic Society.

Bowley, A. 1924. *The mathematical groundwork of economics*. Oxford: Clarendon Press.

Bowley, A. 1928. *Edgeworth's contribution to mathematical statistics*. London: Royal Statistical Society.

Bowley, A. 1934. Francis Ysidro Edgeworth. *Econometrica* 1: 113–124.

Butler, J., and H. Butler. 1927. *The Black Book of Edgeworthtown and other Edgeworth Memories 1585–1817*. London: Faber and Gwyer.

Cournot, A. 1838. *Researches into the mathematical principles of the theory of wealth*, trans. N. Bacon, ed. I. Fisher. London: Stechert-Hafner, 1927.

Creedy, J. 1986. *Edgeworth and the development of neoclassical economics*. Oxford: Basil Blackwell.

Creedy, J. 1988. Wicksell on Edgeworth's tax paradox. *Scandinavian Journal of Economics* 90: 101–112.

Creedy, J. 1990. Marshall and Edgeworth. *Scottish Journal of Political Economy* 37: 18–39.

Creedy, J. 1994a. Exchange equilibria: Bargaining, utilitarian and competitive solutions. *Australian Economic Papers* 33: 34–52.

Creedy, J. 1994b. Launhardt's model of exchange. *Journal of the History of Economic Thought* 16: 40–60.

Edgeworth, M. 1820. *Memories of Richard Lovell Edgeworth Esq., begun by himself and concluded by his daughter, Maria Edgeworth*, 2 vols. London: Hunter.

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk taking. *Journal of Political Economy* 61: 434–435.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Hicks, J., and R. Allen. 1934. A reconsideration of the theory of value. *Economica* 14 (52–76): 196–219.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of demand and supply functions. *Journal of Political Economy* 40: 577–616.

Jenkin, F. 1871. On the principles which regulate the incidence of taxes. Reproduced in *Readings in the economics of taxation*, ed. R. Musgrave and C. Shoup. London: Allen and Unwin, 1959.

Jevons, W. 1871. *The theory of political economy*. 5th ed., ed. H. Jevons. New York: Augustus Kelly, 1957.

Jevons, W. 1881. Review of *Mathematical Psychics*. *Mind* 6: 581–583.

Johnson, W. 1913. The pure theory of utility curves. *Economic Journal* 23: 483–513.

Keynes, J.M. 1972. *Essays in biography. Vol. X of the collected writings of Keynes*. London: Macmillan, for the Royal Economic Society.

Marshall, A. 1890. *Principles of economics*, ed. C. Guillebaud, 2 vols. (variorum ed.). London: Macmillan, 1961.

Marshall, A., and M. Marshall. 1879. *Economics of industry.* London: Macmillan.
McCann, C. 1996. *F.Y. Edgeworth: Writings in probability, statistics and economics*, 3 vols. Cheltenham: Edward Elgar.
Mill, J.S. 1848. *Principles of political economy*. Reprinted with editorial material by W. Ashley. London: Longmans, Green, 1920.
Pigou, A. 1908. *Protective and preferential import duties*. London: Macmillan.
Pigou, A., ed. 1925. *Memorials of Alfred Marshall 1842–1924*. London: Macmillan.
Pigou, A., and D. Robertson. 1931. *Economic essays and addresses*. London: King.
Price, L. 1946. *Memoirs and notes on British economists 1881–1946*. MSS, Brotherton Library, University of Leeds.
Sargan, J. 1976. Econometric estimators and the Edgeworth expansion. *Econometrica* 44: 421–448.
Schumpeter, J. 1954. *History of economic analysis*. London: Allen and Unwin.
Stigler, S. 1978. Francis Ysidro Edgeworth, statistician. *Journal of the Royal Statistical Society, A* 141: 287–322.
Sully, J. 1918. *My life and friends*. London: Fisher Unwin.
Vickrey, W. 1960. Utility, strategy and social decision rules. *Quarterly Journal of Economics* 74: 507–535.
Walras, L. 1874. *Elements of pure economics*. Trans. W. Jaffe. London: Allen and Unwin, 1954.
Whitaker, J., ed. 1975. *The early economic writings of Alfred Marshall 1867–1890*. London: Macmillan.

Edgeworth, Maria (1767–1849)

J. P. Croshaw

Born in England of an Irish land-owning family, Maria Edgeworth began her career as amanuensis and co-author to her father Richard Lovell Edgeworth, the educator and amateur inventor. Her first publications were a series of moral tales for children (*The Parents' Assistant*, 1796, and *Early Lessons*, 1802) which aimed to instil the virtues she saw as essential to a 'good' individual and so a 'good' society: honesty, frugality and hard work. These characteristics match rather precisely those of Adam Smith's 'prudent man' in the *Wealth of Nations.* Her tales teach the value of a work ethic, sharply contrasting the evils of sloth and idleness with the pleasures of diligence and achievement. Indeed, her attitude towards this aspect of labour did not exclude her own privileged class of landowners, who, as she witnessed in her own country, frequently abused the landlord-tenant contract.

In 1800 she published the work which is, perhaps, of most interest to economists, *Castle Rackrent.* Through the character of Thady Quirk, an ancient retainer of the Rackrent family, she recounts the history of three generations of absentee landlords, of their tenants and of the depths to which the Rackrent fortunes had fallen through successive generations of dissolute lifestyle. The book not only influenced prominent literary figures of the time (for example, Turgenev and Walter Scott) but also established a literary precedent for the development of fictional characters within the context of a realistic historical, social and economic setting - an approach which, in England, could be said to reach its peak with George Eliot's *Middlemarch.* In the 19th century the name Rackrent came to stand for the embodiment of the vices of the landed aristocracy and was freely used as such by writers like Carlyle and, later, her nephew F. Y. Edgeworth.

Maria Edgeworth continued her critical examination of the landlord-tenant relationship in novels like *The Absentee* (1812) and *Ennui* (1825) where she addressed issues such as leases, population and economic progress and the impact of manufacture on a traditional agricultural economy. Her letters to David Ricardo confirm her interest in the poverty and distress among the Irish agricultural peasantry. She initiated and engaged in a vigorous correspondence with Ricardo over the potato question and the effects of famines in the 1820s. On this subject she differed with both Ricardo and Malthus arguing that the essential cause of the difficulty lay in mismanagement. She rather amusingly suggested that instead of theorizing from afar, Ricardo should travel to Ireland and see for himself.

Education in Developing Countries

Paul Glewwe

Abstract

In many developing countries, children complete few years of schooling and learn little during their time in school. There are many estimation problems that confound attempts to understand the impact of education policies on years of schooling and learning while in school. Recent research has focused on implementing randomized trials to get around estimation problems based on retrospective data. While some useful results have been found, many additional studies are still under way. As these results accumulate it is likely that general conclusions can be drawn, but the evidence to date is too limited to draw general policy recommendations.

Keywords

Attenuation bias; Credit; Education in developing countries; Education production functions; Returns to schooling; Value of time

JEL Classifications

O1

Most economists who study economic growth agree that an educated citizenry is necessary for sustained economic growth, and virtually all international development organizations concur (UNDP 1990; World Bank 2001), and so those organizations provide substantial financial resources and policy advice to promote education in developing countries. Yet in many developing countries, especially the poorest, many children leave school at a young age and learn little during the time they spend in school. These problems have led many economists and other social scientists to turn their attention to education in developing countries.

This article summarizes recent research on the factors that affect the amount of time that children spend in school and the factors that determine how much they learn during their time in school. Thus, it focuses on the factors that shape education outcomes as opposed to the impact of education on income, economic growth and other phenomena (for a recent assessment of the impact of education on other socioeconomic outcomes, see Glewwe 2002). This article also omits, due to space constraints, a discussion of estimation issues (see Glewwe 2002, and Glewwe and Kremer 2006, for thorough discussions of estimation problems and possible solutions).

Factors that Determine Years of Schooling

In developing countries, parents usually decide how many years their children will attend school. Each year, parents consider the costs and expected benefits of an additional year of schooling and then enrol their children for another year if the expected benefits outweigh the estimated costs. The main costs are school fees and other payments required by schools, transportation and (occasionally) meals and housing, and the opportunity cost of the children's time. There may also be an additional, 'psychic' cost; some parents may dislike particular values that schools attempt to instil in students. For many parents, the largest of these costs is the value of their children's time; in developing countries, especially in rural areas, children's time is valuable because they can help in household farming activities.

The main benefits of schooling are the skills learned (which usually reap substantial monetary returns in the labour market), increased employment opportunities that come with educational credentials, and the direct satisfaction and social approval that parents receive from having educated children. While the decision rule to continue schooling when the benefits outweigh the costs

would seem to hold as a tautology, there are circumstances in which children are not enrolled in school even when the economic benefits outweigh the costs. This could occur because the costs are incurred today while the benefits accrue over many years in the future. In particular, parents who have low incomes and cannot obtain credit may not send their children to school even though the present discounted value at prevailing interest rates is positive.

Given this type of decision making by parents, policies to increase school enrolment must focus on reducing the costs of schooling, increasing the benefits of education, or providing access to credit. Reductions in fees are easy to implement, and in some countries (such as Mexico) parents with low incomes receive monthly payments if their children are enrolled in school. Of course, this entails potentially large budgetary costs, so some governments try to limit fee exemptions and outright subsidies to households or communities that are particularly needy. Evidence from many developing countries indicates that reducing fees or providing payments conditional on school enrolment can lead to large increases in enrolment; studies in Honduras, Kenya, Mexico and Nicaragua document these impacts (see Glewwe and Kremer 2006, for further details and references).

The main alternative policy for increasing school enrolment is to increase the expected returns. These returns will increase if the relative price of skilled labour increases, and if schools become more effective at providing academic skills. While some economists have shown that increased returns to education does raise school enrollment (Foster and Rosenzweig 1996), most policy research has focused on what makes schools more efficient at raising students' skills. This research is discussed in the next section.

Three additional points regarding policies to increase years of schooling deserve attention. First, improvements in the health and nutritional status of both very young and school-age children are another potentially important route to increase the time that children spend in school (see Glewwe and Miguel 2006, for a review of this literature). Second, many policy discussions presume that the main reason children are not in school is that no school is available, yet in most countries schools are available but parents opt not to enrol their children because they judge that the costs outweigh the benefits (see Glewwe and Zhao 2005). Third, the role of credit constraints in determining years in school is an under-researched topic, in terms of both the impact of credit constraints and policies that could loosen those constraints.

Factors that Determine Student Learning

In principle, student learning can be depicted as a production process in which student, household, teacher and school characteristics combine to produce students' academic skills. While the existence of an academic skills production function is true almost by definition, there are serious problems that confound attempts to estimate this process. The main problem is omitted variables bias: students, households, teachers and schools can vary in hundreds of ways, and no data-set contains all variables that are potentially important. Indeed, important factors such as student innate ability, teacher effort and parental encouragement are almost impossible to measure and likely to be correlated with the observed variables. This problem applies to virtually all studies based on retrospective (non-experimental) data; indeed, it is probably the main reason that different studies find very different results (the main alternative explanation is that educational production functions are very different in different countries). A second serious estimation problem is attenuation bias. Much of the data on students, households, teachers and schools has a substantial amount of measurement error. This typically leads to underestimation of the true impacts of variables, which may explain, at least in part, why many variables in estimates of the determinants of student learning are statistically insignificant.

In recent years economists and other social scientists have turned to natural experiments and randomized trials to estimate the impacts of

particular school characteristics, policies and programmes on student academic achievement. Natural experiments result from institutions and policies that cause random variation in school or student characteristics, which can be used to analyse the impact of those characteristics on student learning (and on time spent in school). Randomized trials are controlled experiments designed by researchers and school officials that generate random variation in a school characteristic or policy, which again allows one to estimate the impact of the characteristic or policy on learning. Natural experiments are relatively rare, but in recent years randomized trials have been implemented in many countries in Africa, Asia and Latin America.

One of the first randomized trials was conducted in Nicaragua in the late 1970s. The results indicated that workbooks and radio instruction had significant impacts on pupils' math scores. In the Philippines in the early 1980s, provision of textbooks raised students' performance on academic tests, but in Kenya in the late 1990s the only effect of textbooks was among the better students, perhaps because the textbooks provided were too difficult for most students. Other randomized trials conducted in Kenya suggest little impact on test scores from reductions in class size, provision of flip charts, and provision of deworming medicine. On a more positive note, school meals in Kenya raised test scores in schools that had well-trained teachers, but not in schools with poorly trained teachers. In public schools in an urban area of India, a remedial education programme increased test scores at a relatively low cost. Finally, a computer-assisted learning programme in India also appears to have increased test scores. The positive impacts of radio education in Nicaragua and computer instruction in India suggest that using modern technologies may be particularly helpful in schools with weak teachers. (For citations and more detailed discussion, see Glewwe and Kremer 2006.)

While natural experiments and especially randomized trials may seem to avoid the estimation problems that plague retrospective studies, more randomized studies are needed before general conclusions can be drawn that can guide policy in countries that have not yet had such studies. Moreover, randomized trials can also suffer from estimation problems. One problem is that parents of students in the control schools (or schools excluded from the evaluation) may try to enrol their children in the treatment schools. This may affect the results by increasing class size (if class size affects learning). This would not occur if the policy were implemented nationwide. In addition, children who transfer into treatment schools may not be a random sample of the general student population. A related problem is that marginal students in the treatment schools are less likely to drop out (if the intervention raises student achievement), which leads to underestimation of the impact of the policy on learning if comparisons are made based on all students currently enrolled in school. A final problem with randomized trials is that the evaluation itself may lead the treatment group to change its behaviour, or the control group to change its behaviour, because both groups know that their results are being used in an evaluation.

In summary, recent research on education in developing countries has provided fairly convincing evidence of the impact on time in school and on learning for particular policies in particular countries. Many additional studies are currently under way, and as these results accumulate it is likely that general conclusions can be drawn. This should lead to better education policies, which will contribute to higher economic growth and, ultimately, a higher quality of life in developing countries.

See Also

- ▶ Development Economics
- ▶ Education Production Functions
- ▶ Human Capital
- ▶ Returns to Schooling

Bibliography

Foster, A., and M. Rosenzweig. 1996. Technical change and human capital returns and investments: Evidence from the Green Revolution. *American Economic Review* 86: 931–953.

Glewwe, P. 2002. Schools and skills in developing countries: Education policies and socioeconomic outcomes. *Journal of Economic Literature* 40: 436–482.
Glewwe, P., and M. Kremer. 2006. Schools, teachers and education outcomes in developing countries. In *Handbook on the economics of education*, ed. E. Hanushek and F. Welch. Amsterdam: North-Holland.
Glewwe, P., and E. Miguel. 2006. The impact of child health and nutrition on education in less developed countries. In *Handbook of agricultural economics*, vol. 4, ed. R. Evenson and T. Schultz. Amsterdam: North-Holland.
Glewwe, P., and M. Zhao. 2005. Attaining universal primary completion by 2015: How much will it cost?. Department of Applied Economics/University of Minnesota.
UNDP (United Nations Development Programme). 1990. *Human development report*. New York: UNDP.
World Bank. 2001. *World development report 2000/2001: Attacking poverty*. Washington, DC: World Bank.

Education Production Functions

Eric A. Hanushek

Abstract

The accumulated economic analysis of education suggests that current provision of schooling is very inefficient. Commonly purchased inputs to schools – class size, teacher experience, and teacher education – bear little systematic relationship to student outcomes, implying that conventional input policies are unlikely to improve achievement. At the same time, differences in teacher quality have been shown to be very important. Unfortunately, teacher quality, defined in terms of effects on student performance, is not closely related to salaries or readily identified attributes of teachers.

Keywords

Education production functions; Random assignment; School attainment; School resources; Student outcomes; Teacher quality

JEL Classifications

I2

A simple production model lies behind much of the analysis in the economics of education. The common inputs are things like school resources, teacher quality, and family attributes; and the outcome is student achievement. Knowledge of the production function for schools can be used to assess policy alternatives and to judge the effectiveness and efficiency of public provided services. This area is, however, distinguished from many because the results of analyses enter quite directly into the policy process.

Historically, the most frequently employed measure of schooling has been attainment, or simply years of schooling completed. The value of school attainment as a rough measure of individual skill has been verified by a wide variety of studies of labour market outcomes (for example, Mincer 1970; Psacharopoulos and Patrinos 2004). However, the difficulty with this common measure of outcomes is that it assumes a year of schooling produces the same amount of student achievement, or skills, over time and in every country. This measure simply counts the time spent in schools without judging what happens in schools – thus, it does not provide a complete or accurate picture of outcomes.

Recent direct investigations of cognitive achievement find significant labour market returns to individual differences in cognitive achievement (for example, Lazear 2003; Mulligan 1999; Murnane et al. 2000). Similarly, society appears to gain in terms of productivity; Hanushek and Kimko (2000) demonstrate that quality differences in schools have a dramatic impact on productivity and national growth rates. (A parallel line of research has employed school inputs to measure quality but has not been as successful. Specifically, school input measures have not proved to be good predictors of wages or growth.)

Because outcomes cannot be changed by fiat, much attention has been directed at inputs – particularly those perceived to be relevant for policy such as school resources or aspects of teachers.

Analysis of the role of school resources in determining achievement begins with the Coleman Report, the US government's monumental study on educational opportunity released

in 1966 (Coleman et al. 1966). That study's greatest contribution was directing attention to the distribution of student performance – the outputs as opposed to the inputs.

The underlying model that has evolved as a result of this research is very straightforward. The output of the educational process – the achievement of individual students – is directly related to inputs that both are directly controlled by policymakers (for example, the characteristics of schools, teachers, and curricula) and are not so controlled (such as families and friends and the innate endowments or learning capacities of the students). Further, while achievement may be measured at discrete points in time, the educational process is cumulative; inputs applied sometime in the past affect students' current levels of achievement.

Family background is usually characterized by such socio-demographic characteristics as parental education, income, and family size. Peer inputs, when included, are typically aggregates of student socio-demographic characteristics or achievement for a school or classroom. School inputs typically include teacher background (education level, experience, sex, race, and so forth), school organization (class sizes, facilities, administrative expenditures, and so forth), and district or community factors (for example, average expenditure levels). Except for the original Coleman Report, most empirical work has relied on data constructed for other purposes, such as a school's standard administrative records. Based upon this, statistical analysis (typically some form of regression analysis) is employed to infer what specifically determines achievement and what is the importance of the various inputs into student performance.

Measured School Inputs

The state of knowledge about the impacts of resources is best summarized by reviewing available empirical studies. Most analyses of education production functions have directed their attention at a relatively small set of resource measures, and this makes it easy to summarize the results (Hanushek 2003). The 90 individual publications that appeared before 1995 contain 377 separate production function estimates. For classroom resources, only nine per cent of estimates for teacher education and 14% for teacher–pupil ratios yielded a positive and statistically significant relationship between these factors and student performance. Moreover, these studies were offset by another set of studies that found a similarly negative correlation between those inputs and student achievement. Twenty-nine per cent of the studies found a positive correlation between teacher experience and student performance; however, 71% still provided no support for increasing teacher experience (being either negative or statistically insignificant). Studies on the effect of financial resources provide a similar picture. These indicate that there is very weak support for the notion that simply providing higher teacher salaries or greater overall spending will lead to improved student performance. Per pupil expenditure has received the most attention, but only 27% of studies showed a positive and significant effect. In fact, seven per cent even suggested that adding resources would harm student achievement. It is also important to note that studies involving pupil spending have tended to be the lowest-quality studies as defined below, and thus there is substantial reason to believe that even the 27% figure overstates the true effect of added expenditure.

These studies make a clear case that resource usage in schools is subject to considerable inefficiency, because schools systematically pay for inputs that are not consistently related to outputs.

Study Quality

The previous discussions do not distinguish among studies on the basis of any quality differences. The available estimates can be categorized by a few objective components of quality. First, while education is cumulative, frequently only current input measures are available, which results in analytical errors. Second, schools operate within a policy environment set almost always at higher levels of government. In the United

States, state governments establish curricula, provide sources of funding, govern labour laws, determine rules for the certification and hiring of teachers, and the like. In other parts of the world, similar policy setting, frequently at the national level, affects the operations of schools. If these attributes are important – as much policy debate would suggest – they must be incorporated into any analysis of performance. The adequacy of dealing with these problems is a simple index of study quality.

The details of these quality issues and approaches for dealing with them are discussed in detail elsewhere (Hanushek 2003) and only summarized here. The first problem is ameliorated if one uses the 'value added' versus 'level' form in estimation. That is, if the achievement relationship holds at different points in time, it is possible to concentrate on the growth in achievement and on exactly what happens educationally between those points when outcomes are measured. This approach ameliorates problems of omitting prior inputs of schools and families, because they will be incorporated in the initial achievement levels that are measured (Hanushek 1979). The latter problem of imprecise measurement of the policy environment can frequently be ameliorated by studying performance of schools operating within a consistent set of policies – for example, within individual states in the USA or similar decision-making spheres elsewhere. Because all schools within a state operate within the same basic policy environment, comparisons of their performance are not strongly affected by unmeasured policies (Hanushek et al. 1996).

If the available studies are classified by whether or not they deal with these major quality issues, the prior conclusions about research usage are unchanged (Hanushek 2003). The best quality studies indicate no consistent relationship between resources and student outcomes.

An additional issue, which is particularly important for policy purposes, concerns whether this analytical approach accurately assesses the causal relationship between resources and performance. If, for example, school decision-makers provide more resources to those they judge as most needy, higher resources could simply signal students known for having lower achievement. Ways of dealing with this include various regression discontinuity or panel data approaches. When done in the case of class sizes, the evidence has been mixed (Angrist and Lavy 1999; Rivkin et al. 2005).

An alternative involves the use of random assignment experimentation rather than statistical analysis to break the influence of sample selection and other possible omitted factors. With one major exception, this approach nonetheless has not been applied to understand the impact of schools on student performance. The exception is Project STAR, an experimental reduction in class sizes that was conducted in the US state of Tennessee in the mid-1980s (Word et al. 1990). To date, it has not had much impact on research or our state of knowledge. While Project STAR has entered into a number of policy debates, the interpretation of the results remains controversial (Krueger 1999; Hanushek 1999).

Magnitude of Effects

Throughout most consideration of the impact of school resources, attention has focused almost exclusively on whether a factor has an effect on outcomes that is statistically different from zero. Of course, any policy consideration would also consider the magnitude of the impacts and where policies are most effective. Here, even the most refined estimates of, say, class size impacts does not give very clear guidance. The experimental effects from Project STAR indicate that average achievement from a reduction of eight students in a classroom would increase by about 0.2 standard deviations, but only in the first grade of attendance in smaller classes (kindergarten or first grade) (see Word et al. 1990; Krueger 1999). Angrist and Lavy (1999), with their regression discontinuity estimation, find slightly smaller effects in grade five and approximately half the effect size in grade four. Rivkin et al. (2005), with their fixed effects estimation, find effects half of Project STAR in grade four and declining to insignificance by grade seven. Thus, from a policy perspective the alternative estimates are both small in economic

terms when contrasted with the costs of such large class size reductions and inconsistent across studies.

Do Teachers and Schools Matter?

Because of the Coleman Report and subsequent studies discussed above, many have argued that schools do not matter and that only families and peers affect performance. Unfortunately, these interpretations have confused measurability with true effects.

Extensive research since the Coleman Report has made it clear that teachers do indeed matter when assessed in terms of student performance instead of the more typical input measures based on characteristics of the teacher and school. When fixed effect estimators that compare student gains across teachers are used, dramatic differences in teacher quality are seen.

These results can also be reconciled with the prior ones. These differences among teachers are simply not closely correlated with commonly measured teacher characteristics (Hanushek 1992; Rivkin et al. 2005). Moreover, teacher credentials and teacher training do not make a consistent difference when assessed against student achievement gains (Boyd et al. 2006; Kane et al. 2006). Finally, teacher quality does not appear to be closely related to salaries or to market decisions. In particular, teachers exiting for other schools or for jobs outside of teaching do not appear to be of higher quality than those who stay (Hanushek et al. 2005).

Some Conclusions and Implications

The existing research suggests inefficiency in the provision of schooling. It does not indicate that schools do not matter. Nor does it indicate that money and resources never impact achievement. The accumulated research surrounding estimation of education production functions simply says there currently is no clear, systematic relationship between resources and student outcomes.

See Also

- ▶ Human Capital
- ▶ Local Public Finance
- ▶ Returns to Schooling

Bibliography

Angrist, J.D., and V. Lavy. 1999. Using Maimondides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114: 533–575.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy* 1: 176–216.

Coleman, J.S., E.Q. Campbell, C.J. Hobson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York. 1966. *Equality of educational opportunity*. Washington, DC: US Government Printing Office.

Hanushek, E.A. 1979. Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources* 14: 351–388.

Hanushek, E.A. 1992. The trade-off between child quantity and quality. *Journal of Political Economy* 100: 84–117.

Hanushek, E.A. 1999. Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis* 21: 143–163.

Hanushek, E.A. 2003. The failure of input-based schooling policies. *Economic Journal* 113: F64–F98.

Hanushek, E.A., and D.D. Kimko. 2000. Schooling, labor force quality, and the growth of nations. *American Economic Review* 90: 1184–1208.

Hanushek, E.A., S.G. Rivkin, and L.L. Taylor. 1996. Aggregation and the estimated effects of school resources. *Review of Economics and Statistics* 78: 611–627.

Hanushek, E.A., J.F. Kain, D.M. O'Brien, and S.G. Rivkin. 2005. *The market for teacher quality*, Working paper no. 11154. Cambridge, MA: NBER.

Kane, T.J., J.E. Rockoff, and D.O. Staiger. 2006. *What does certification tell us about teacher effectiveness? Evidence from New York City*, Working paper no. 12155. Cambridge, MA: NBER.

Krueger, A.B.. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114: 497–532.

Lazear, E.P. 2003. Teacher incentives. *Swedish Economic Policy Review* 10(3): 179–214.

Mincer, J. 1970. The distribution of labor incomes: A survey with special reference to the human capital approach. *Journal of Economic Literature* 8: 1–26.

Mulligan, C.B. 1999. Galton versus the human capital approach to inheritance. *Journal of Political Economy* 107(pt. 2): S184–S224.

Murnane, R.J., J.B. Willett, Y. Duhaldeborde, and J.H. Tyler. 2000. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management* 19: 547–568.

Psacharopoulos, G., and H.A. Patrinos. 2004. Returns to investment in education: A further update. *Education Economics* 12: 111–134.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73: 417–458.

Word, E., J. Johnston, H.P. Bain, B. DeWayne Fulton, J.B. Zaharies, M.N. Lintz, C.M. Achilles, J. Folger, and C. Breda. 1990. *Student/Teacher Achievement Ratio (STAR), Tennessee's K-3 class size study: Final summary report, 1985–1990*. Nashville: Tennessee State Department of Education.

Educational Finance

William A. Fischel

Abstract

The American system of government-financed education is decentralized among 50 states and more than 15,000 local school districts. Local funds are derived from local property taxes, and this system tends to make local spending unequal. State- government efforts to equalize education spending involve manipulating the local 'tax price' with matching grants. School districts with low tax prices are not, however, necessarily populated by rich people, so the distribution of state funds may penalize many low-income districts with large amounts of non-residential property.

Keywords

Educational finance; Local government; Median voter; Property taxation; School districts (USA); School vouchers; Spatial competition; Tax price of school spending; Tiebout hypothesis

JEL Classification

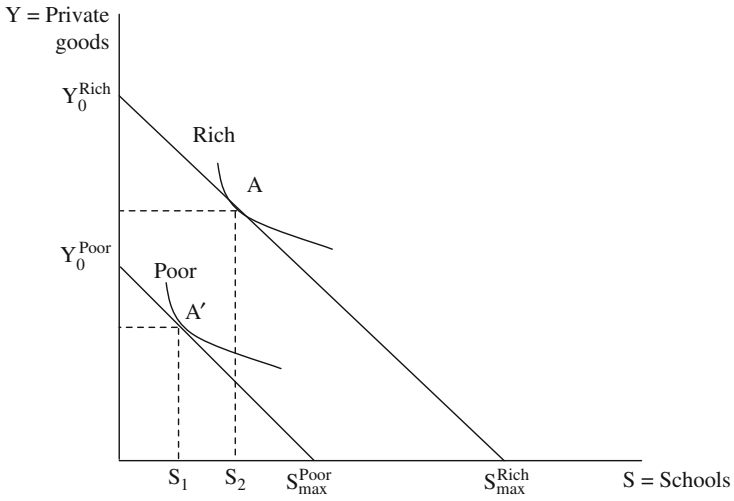
I2

This article deals with the government-financed system of education in the United States, which is referred to as 'public' education. Educational finance in the United States is different from that of other nations, which typically fund education from national taxes. Within each American state, a substantial portion of education is financed by local governments, although the proportion financed locally has declined from 83.2% in 1920 to 43.2% in 2000.

The state–local system of finance stems from the history and geography of the United States and the federal nature of its government. The 50 states are, in the eyes of the national government, primarily responsible for education. In most states, implementation of this responsibility is delegated to local municipal corporations called 'school districts'. The school district is more than a local administrative agency of the state. It is a distinct political entity that usually has some correspondence with the geographic area of a municipality. The district, however, has a separate board of directors, which is locally elected. The board then selects a superintendent of schools to manage the district's education. Boards have the authority to levy taxes, which are almost always on property within their district, and spend the revenue they derive from them. The state government may prescribe curricular standards for public schools, but the method of achieving these standards is the responsibility of the local district.

School districts and school boards were once the most common form of local government in the United States, numbering about 200,000 in 1900. The number of school districts declined steadily throughout the twentieth century, which can largely be accounted for by the consolidation of rural one-room school districts into larger units. By 1970, one-room schools were essentially extinct, and since 1970 the total number of school districts has declined only slightly, numbering about 16,000 at the beginning of the twenty-first century.

Despite their numerical decline in rural areas, there are many school districts in most metropolitan areas. Urban households that are already on the move for job- related reasons have the luxury of choosing a home within one of several school



Educational Finance, Fig. 1 School spending in rich and poor districts

districts in most regions of the nation. Choosing among school districts and the resulting competition among districts to obtain residents is consistent with the model proposed by Tiebout (1956). Numerous tests of the Tiebout model indicate that the quality of schooling is important to most home buyers (Oates 1969; Bradbury et al. 2001). There is also evidence that spatial competition makes school districts more efficient in delivering education services (Hoxby 2000).

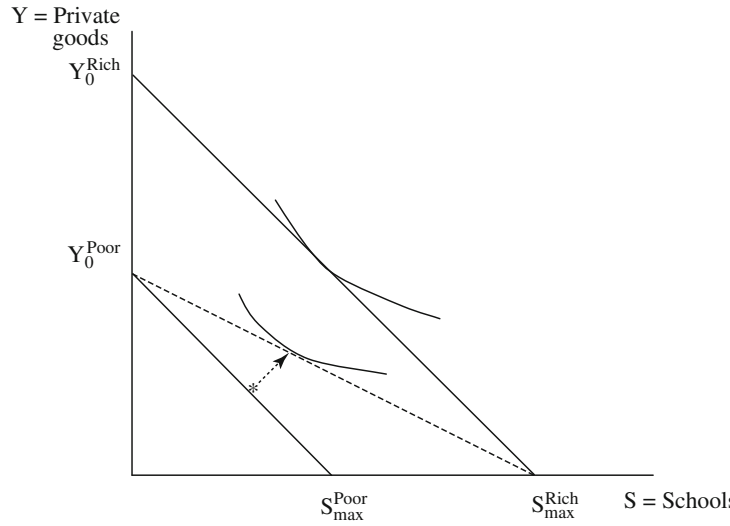
One-room schools of the nineteenth century were usually 'ungraded'. Students were instead divided into skill-specific recitation groups, formed without regard for chronological age. In this system, uniformity of education was not critical. New pupils could be placed according to what they knew in particular subjects rather than by age. But when almost all schools were age-graded, it paid for each district to offer an age-specific curriculum that allowed both teachers and pupils to be interchangeable among schools and districts (Fischel 2006a).

Standardization of age-graded curricula became widespread by about 1940 and was brought about by two forces, one local and the other statewide. Property-owning voters in a given district would find that potential homebuyers would shun them if they did not offer a standard, public-school education. Voters would thus support taxes necessary to fund standardized schools. However, differences in the economic make-up and tax-bases of local districts sometimes made this difficult to do.

Figure 1 illustrates the problem for attempts to fund schools from local sources. It depicts a trade-off between local school spending and other goods for the median voter (the voter with the median income, assumed always to be in the majority in local elections) in two separate communities, a rich district and a poor district. The decisive voter chooses the mix of school spending and private goods that achieves the highest indifference curve that his private–public budget line allows (Bergstrom and Goodman 1973). Because at the local level education is essentially a private good, the slope of the budget lines is the 'tax price' of school spending for the median voter in each community.

The tax price is not a tax rate. A school district composed exclusively of mansions will have, for a given level of spending, a much lower property tax *rate* than a district composed of modest-sized homes. But if the second moment of the distribution of wealth is the same in both communities, the *tax price* faced by the median voter in each will be the same. A 1000 dollar increase in per-pupil spending will cost the median voter the same amount of money in both cases, if one assumes that the number of public-school children per household is the same in both.

The other generalization that Fig. 1 illustrates is that average income of a district accounts for



Educational Finance, Fig. 2 Subsidies to poor districts

much of the differences in spending per pupil. Even though the tax prices are the same, the positive income elasticity of demand for education (estimated at somewhere between 0.5 and 1.0) causes the richer community to choose a higher level of school inputs (Bergstrom et al. 1982). While much of the criticism of these differences is based on equity concerns, there are efficiency reasons to promote a relatively uniform system of education (Benabou 1996).

The way most states have attempted to equalize education opportunities is to reduce the tax price of spending in poorer districts. State funds (from statewide taxes) are offered to the poorer community in proportion to the district's own tax effort. The poorer median voter thus perceives, as indicated by the dotted budget line in Fig. 2, that for every dollar raised locally, the state will send it another dollar. The tax price has been cut in half in the graphical example, so that the poorer community will choose to spend an amount closer to that of the richer district.

By manipulating the local tax price, state governments can in principle induce a substantial equality of school spending in nominally independent districts, though state officials still seem surprised that there is an income effect as well as a substitution effect from lowering the tax price. They seem to expect that the arrow in Fig. 2 should point horizontally to the right. Instead, local voters use the subsidy (the reduced tax price) to both increase local spending on schools, which is the desired substitution effect, and to reduce their own local taxes (nudging the arrow's direction upwards), which is the income effect.

Another factor can also account for differences in local tax prices. The poorer district may have a substantial amount of non-residential property to tax. Commercial and industrial uses do not come with children attached (at least in metropolitan areas, where workers can live in other communities), and so their tax revenues amount to a subsidy to their school district. The effect of this is the same as a matching-grant subsidy by the state. And the effect is not trivial. Nationally, almost one-half of all property taxes are paid by non-residential property owners, which puts them on the same order of magnitude as state funds for public education.

Although both state subsidies and a large non-residential tax base reduce the tax price, they have been treated differently in recent years. The school finance litigation movement began with *Serrano v. Priest* in California in 1971 (Brunner and Sonstelie 2006). Its objective was to use state constitutional directives (equal protection and school funding clauses) to improve schools in poor districts. For strategic reasons, the movement focused its remedial efforts on

differences in tax base per pupil rather than differences in spending per pupil or on educational outcomes. Many state courts thus ruled that unequal tax bases, not unequal spending, were constitutionally suspect and ordered legislatures to transfer funds from the 'property rich' to the 'property poor'.

What this remedy overlooked is that low-income communities are as likely to be 'property rich' (on the widely used 'tax base per-pupil' standard) as high-income communities. This is because many urban districts have a large non-residential property tax base that offsets the lower valued residential tax base. (The poor may have migrated there for jobs or rezoned land to attract industry, something most affluent suburbs are reluctant to do.) Besides this, poorer cities often have relatively few children in public schools because of an aged population or because low-quality public schools encourage the use of private schools. In any case, many of the court-induced 'equalization' remedies have actually caused state funds to be removed from low-income (but 'property rich') districts to higher-income districts that are 'property poor' because of their modest nonresidential tax base and large school-age population.

An alternative response to the difficulties of distributing state funds to school districts is simply to have the state government run the schools without the intermediation of local school boards and districts. Another is a voucher system, in which the state gives public funds to parents and allows them to select whatever school they want. Both are certainly viable means of school finance, and it is worth asking why they have not been embraced.

Full state funding forgoes the local monitoring of school performance by voters. Capitalization of school quality in local home values creates a feedback mechanism for local governance. The median voter in most jurisdictions is a homeowner, and voters therefore care about the consequences of school governance. School superintendents who waste local taxpayers' money will find that their tenure is short as voters become dissatisfied. Even if they keep their jobs, the declines in taxable property value due to inefficient policies will leave them with less revenue to spend in the future (Hoxby 1999). Neither of these desirable feedback effects is likely to occur under a state-managed system.

The drawback of school vouchers appears to be that voters are reluctant to embrace them as a general practice. American voters appear to perceive benefits from local public schools that go beyond educational qualities. One benefit I have advanced is that public schools create location-specific social capital among adults (Fischel 2006b). Adults with children are more likely to know the parents of their children's schoolmates. This creates a network of adult social capital that lowers the transaction costs of public participation in municipal affairs. A voucher system disperses children to various schools and thus does not create the same location- specific social capital that public schools do. In any case, America's continuing embrace of locally run and locally financed public education reflects the school's central role in facilitating local self-governance.

See Also

- Exit and Voice
- Fiscal Federalism
- Local Public Finance
- Property Taxation
- Public Choice
- School Choice and Competition
- Tiebout Hypothesis

Bibliography

Benabou, R. 1996. Heterogeneity, stratification, and growth: Macroeconomic implications of community structure and school finance. *American Economic Review* 86: 584–609.

Bergstrom, T.C., and R.P. Goodman. 1973. Private demand for public goods. *American Economic Review* 63: 280–296.

Bergstrom, T.C., D.L. Rubinfeld, and P. Shapiro. 1982. Micro-based estimates of demand functions for local school expenditures. *Econometrica* 50: 1183–1205.

Bradbury, K.L., K.E. Case, and C. Mayer. 2001. Property tax limits, local fiscal behavior, and property values: Evidence from Massachusetts under proposition 2 1/2. *Journal of Public Economics* 80: 287–311.

Brunner, E.J., and J. Sonstelie. 2006. California's school finance reform: An experiment in fiscal federalism. In *The Tiebout model at fifty*, ed. W.A. Fischel. Cambridge, MA: Lincoln Institute of Land Policy.
Fischel, W.A. 2006a. Will I see you in September? An economic explanation for the standard school calendar. *Journal of Urban Economics* 59: 236–251.
Fischel, W.A. 2006b. Why voters veto vouchers: Public schools and community- specific social capital. *Economics of Governance* 7: 109–132.
Hoxby, C.M. 1999. The productivity of schools and other local public goods producers. *Journal of Public Economics* 74: 1–30.
Hoxby, C.M. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90: 1209–1238.
Oates, W.E. 1969. The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy* 77: 957–971.
Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Effective Demand

J. A. Kregel

Abstract

By 'effective demand' Keynes meant the forces determining changes in the scale of output and employment as a whole. It was intended to replace Say's Law. For Keynes, since entrepreneurs maximized monetary returns, not employment or physical output, there was no reason why their investment decisions should lead to an equilibrium at full employment. Since this account permitted any level of employment to emerge as a stable equilibrium, including full employment, it is more general than the classical Say's Law position, in which the only stable equilibrium was the limit set by full employment as given in the labour market.

Keywords

Effective demand; Excess supply; Full employment equilibrium; Gold; Hume, D.; Liquidity preference; Malthus, T. R.; Marginal efficiency of capital; Mercantilism; Mill, J. S.; Multiplier; Natural price; Natural rate and market rate of interest; Propensity to consume; Quantity theory of money; Rate of interest; Saving and investment; Say's Law; Smith, A.; Subjective theory of value; Supply and demand; Velocity of circulation

JEL Classifications

E12

'Effective demand' is the term used by Keynes in his *General Theory* (1936a) to represent the forces determining changes in the scale of output and employment as a whole. Keynes attributed the first discussions of the determinants of the supply and demand for output as a whole to the classical economists, in particular the debate between Ricardo and Malthus concerning the possibility of 'general gluts' of commodities, or what has come to be known as Say's Law of Markets. Indeed, Keynes's theory was intended to replace Say's Law, although the emergence of effective demand from his *Treatise on Money* (1930) critique of the quantity theory of money, and his insistence on its application in what he originally called a 'monetary production economy', suggests that it should also be seen in antithesis to classical monetary theory. For Adam Smith (1776, p. 285), 'A man must be perfectly crazy who . . . does not employ all the stock which he commands, whether it be his own or other peoples' on consumption or investment. As long as there was what Smith called 'tolerable security', economic rationality implied that it was impossible for demand for output as a whole to diverge from aggregate supply. Although Smith (1776, p. 73) did call the demand 'sufficient to effectuate the bringing of the commodity to the market', the 'effectual demand' 'of those who are willing to pay the natural price' of the commodity, the idea referred to divergence of market from natural price of particular commodities and the process of gravitation of prices to their natural values. J.B. Say's discussion of the problem of the 'disposal of commodities' adopted Smith's position. Against those who held that 'products would

always be abundant, if there were but a ready demand, or market for them,' Say's 'law of markets' argued 'that it is production which opens a demand for products' (1855, pp. 132–3); if production determined ability to buy, then demand could not be deficient. While excesses in particular markets were admitted, they would always be offset by deficiencies in others. Ricardo used similar arguments against Malthus, who responded by suggesting that:

> from the want of a proper distribution of the actual produce, adequate motives are not furnished to continued production, . . . the grand question is whether it [actual produce] is distributed in such a manner between the different parties concerned as to occasion the most effective demand for future produce . . . (Malthus 1821)

Malthus argues that the composition of output affects its quantity by producing doubts in the minds of Smith's rational entrepreneurs concerning the 'security' of their future profit.

The final word in the classical debate was J.S. Mill's 'On the Influence of Consumption on Production', which sought exceptions to the proposition that 'All of which is produced is already consumed, either for the purpose of reproduction or enjoyment' so that 'There will never, therefore, be a greater quantity produced, of commodities in general, than there are customers for' (1874, pp. 48–9). Mill accused those who argued that demand limits output of a fallacy of composition, for the individual shopkeeper's failure to sell is due to a disproportion of demand which cancels out for the nation as a whole. Mill also notes that the argument that every purchaser must be a seller presumes barter, for money enables exchange ' to be divided into two separate acts' so one 'need not buy at the same moment when he sells' (p. 70). To avoid this problem 'money must itself be considered as a commodity', for 'there cannot be an excess of all other commodities, and an excess of money at the same time' (p. 71). Mill admits that if money were 'collected in masses', there might be an excess of all commodities, but this would mean only a temporary fall in the value of all commodities relative to money. Similarly to Smith's 'tolerable security', Mill explains an excess of commodities in general by 'a want of commercial confidence', which he denies may be caused by an overproduction of commodities (p. 74).

Mill's defence of Say's Law highlights the importance of the classical quantity theory, which was originally formulated to oppose the undue emphasis given to precious metals as components of national wealth by the mercantilists. Hume noted that labour, not gold, produced the commodities which composed national wealth; that gold was only as good as the labour it commanded to produce output. Thus the classical position that the velocity of circulation of money was independent of its quantity was built on the view that money would only be held to be spent. Money could at best cause temporary general gluts; in the long term, 'rational' men would not choose to hold money rather than spend it.

On the eve of the marginal revolution, classical theory thus admitted the temporary occurrence of general gluts explained by cyclical disproportions in demand for money and commodities due to crises of confidence. It is paradoxical that, while the marginal revolution was motivated by the failure of classical theory to give sufficient attention to the role of demand in value theory, it failed to extend its analysis of demand to output as a whole in either the long or the short period. Indeed, the emphasis on individual equilibrium produced by the subjective theory of value which replaced the classical theory, made separate discussion of aggregate supply and demand redundant. Thus Keynes's reference to 'the disappearance of the theory of demand and supply for output as a whole, that is the theory of employment *after* it has been for a quarter of a century the most discussed thing in economics' (Keynes 1936c).

But it was discussion, not Say's Law, which disappeared from neoclassical economics. Thus Keynes classed economists from Smith and Ricardo to Marshall and Pigou as 'Classical', for, despite antagonistic theories of value and distribution, they all held a similar theory of supply and demand for output as a whole.

Keynes suggests that this was due more to the failure of neoclassical economists to heed Mill's warning concerning the extension of the

conditions faced by the individual to the economy as a whole, than to positive analysis. If consumers (producers) maximize utility (profit) subject to an income (cost) constraint, reaching the maximum by substituting in consumption (production) goods (inputs) which were cheaper per unit of utility (output), then excess supply of any good (resource) is due to its price exceeding its marginal utility (productivity). Market competition would lead to relative price adjustments which eliminate excess supply. Since it was impossible for any single good (resource) to be unsold (unemployed), it was natural to extend this analysis to the aggregate level to deny the possibility of general gluts without further analysis.

Any divergence from this position was explained, not by reference to hoarding money due to crises of confidence, but by temporary impediments to the automatic adjustment of relative prices in competitive markets. Thus, despite their new marginal theory of value, Keynes's contemporaries reached a similar result that divergence of employment from its full employment level would be determined by temporary non-persistent causes eliminated in the long run.

From 1921 to 1939 the unemployment rate in the United Kingdom never fell below ten per cent, peaking in 1932 at 22.5 per cent (over 2.7 million). This exceeded the limits that most economists attributed to short-period frictions. The self-adjusting nature of the neoclassical version of Say's Law that Keynes chose to criticize was thus contradicted by reference to economic events as well as by Keynes's conception of effective demand.

Keynes was not concerned with impediments to the equality of the supply and demand, but with the

> problem of the equilibrium of supply and demand for output as a whole, in short, of effective demand . . . When one is trying to discover the volume of output and employment, it must be this point of equilibrium for which one is searching.

While the Classics solved the problem by assuming the identity of savings and expenditure on investment goods, neoclassical theory presumed Say's Law 'without giving the matter the slightest discussion' (1936b, p. 215).

Keynes's theory of effective demand thus had to replace Say's Law. To do this Keynes departed from the Classical position on two points. The first was to assume that wages exceed subsistence so that expenditure on consumption goods does not exhaust factor incomes. As expressed in Keynes's psychological law of consumption, this implied that as output increased, the gap between aggregate expenditure and factor costs increased, so that unless investment expenditure expanded to fill the gap, entrepreneurs would experience losses.

The second departure was from the assumption that rationality dictated that entrepreneurs' savings represented productive investment expenditure. If investment could produce losses, or changes in interest rates change capital values, then greater future enjoyment might be assured by not investing; holding money might be 'rational' in such conditions. Further, in a monetary economy, nothing guarantees that maximization of returns in money will maximize either productive capacity or the demand for labour.

In Keynes's theory the propensity to consume and the multiplier produce the proposition that it is the level of output which adjusts saving to investment, rather than the rate of interest, while the explanation of the decisions over the level of investment in a monetary economy requires an explanation of rates of interest in money terms. The two factors are closely related.

In a 1934 letter to Kahn, Keynes gives a 'precise definition of what is meant by effective demand' (1934a, p. 422). If O is the level of output, W the marginal prime cost of production for that output, and P the expected selling price, 'Then OP is effective demand'. The classical theory that 'supply creates its own demand' assumes that OP equals OW, irrespective of the value of O, 'so that effective demand is incapable of setting a limit to employment which consequently depends on the relation between marginal product in wage-goods industries and marginal disutility of employment'. Thus, what Keynes later called (1936a, ch. 2) the two 'classical' postulates limit O at full employment. In contrast,

> On my theory OW $\neq$ OP for *all* values of O, and entrepreneurs have to choose a value of O for which it is equal – otherwise the equality of price and

> marginal prime cost is infringed. This is the real starting point of everything.

The key point was thus the impact of different levels of O on the difference between costs and prices, that is on entrepreneurs' profits. Keynes took up this question, in an undated exchange with Sraffa of about the same time (1934b, pp. 157ff). Keynes notes that a non-unitary marginal propensity to consume implies OP $\neq$ OW for any O, and generates.

> the general principle that *any* expansion of output gluts the market unless there is a *pari passu* increase of investment appropriate to the community's marginal propensity to consume; and any contraction leads to windfall profits to producers unless there is an appropriate *pari passu* contraction of investment.

The level of O at which OP = OW will be determined by the level of investment and the propensity to consume. Changes in the rate of investment, based on entrepreneurs' expectations of their future profits, will determine O.

In an early draft of the *General Theory* Keynes (1973a, p. 439) put it this way:

> Effective demand is made up of the sum of two factors based respectively on the expectation of what is going to be consumed and on the expectation of what is going to be invested.

Thus the theory of effective demand required, in addition to explanation of consumption based on the propensity to consume, an explanation of variations in the level of investment. Since neoclassical theory resolved this problem by presuming that investment was brought into balance with full employment saving by means of the rate of interest, Keynes located the 'flaw being largely due to the failure of the Classical doctrine to develop a satisfactory theory of the rate of interest' (1934c, p. 489).

Keynes concentrated his efforts to produce a theory of interest compatible within this theory of effective demand within what he called a monetary production economy. The *Treatise on Money* (1930) had explained changes in prices in terms of households' consumption decisions relative to entrepreneurs' production decisions. If these decisions were incompatible, investment diverged from saving and prices of consumption goods adjusted producing windfall profits or losses. The prices of investment goods were determined separately from this process, by means of the interaction of the bearishness of the public reflecting their decisions to hold bank deposits or securities on the one hand, and the monetary policy of the banking system on the other.

Investment goods are held because their present costs or supply prices are lower than the present value of their anticipated future earnings or demand prices; the larger this difference, the higher the expected rate of return. Since any change in the price of a durable capital asset will influence its rate of return, a theory that explains the price of capital assets also explains rates of return (which Keynes called marginal efficiency). With the demand price of an asset based on the value of expected future earnings discounted by the rate of interest, it is clear why a satisfactory theory of interest is crucial to the explanation of effective demand.

But money was a durable asset like any other, and as such it has a spot or demand price and a supply price or forward price, which determine the money rate of interest. Keynes thus transformed his concept of bearishness into liquidity preference which, together with banking policy, would determine the rate of interest. For Keynes, 'the money rate of interest . . . is nothing more than the percentage excess of a sum of money contracted for forward delivery . . . over what we may call the "spot" or cash price of the sum thus contracted for forward delivery' (1936a, p. 222), it is:

> the premium obtainable on current cash over deferred cash . . . No one would pay this premium unless the possession of cash served some purpose, that is had some efficiency. Thus we may conveniently say that interest on money measures the marginal efficiency of money measured in terms of itself as a unit. (1937a, p. 101)

Since both money and capital assets had marginal efficiencies representing their rates of return, profit-maximizing individuals in a monetary economy would demand money and capital assets in proportions which equated their respective returns. The equilibrium level of output chosen by entrepreneurs would then be represented by

equality of the marginal efficiency of capital and the rate of interest (the marginal efficiency of money). The question of the effect of an increase in output on profit raised by a propensity to consume less than unity can now be seen as the effect of an increase in investment on the marginal efficiency of money relative to the marginal efficiencies of capital assets. Since these marginal efficiencies reflect pairs of spot and forward asset prices, the question can also be put as the effect of an increase in investment on relative money prices. Thus Keynes's independent variables, the propensity to consume, the efficiency of capital and liquidity preference, given expectations and monetary policy, interact to determine effective demand.

Since this equilibrium could be described by S = I, or equality between the rate of interest and the marginal efficiency of capital, the level of output which equates aggregate demand and supply also equates marginal efficiency with the rate of interest. To complete his theory of effective demand, Keynes faced the question first raised by Wicksell of the causal relation between the natural and the money rate of interest. Just as Keynes rejected the determination of the level of O at which OP = OW by the equality of the marginal productivity and disutility of labour, he rejected marginal productivity as the determinant of marginal efficiency and the real rate of interest determining the money rate because it was based on 'circular reasoning' (1937b, p. 212).

Keynes argues instead that it is the marginal efficiency of capital assets which adapts to the money rate of interest rather than vice versa. These two points of departure are discussed in Chapters 16 and 17 of the *General Theory,* where Keynes points out that the money rate of return to be expected from a capital asset depends on the relation of anticipated money receipts relative to expected money costs, and that there is no reason to believe that these will be related in any predictable way to the asset's physical productivity. Wicksell's natural rate, derived from physical relations of production and exchange, has no application in a monetary economy; Keynes thus substitutes the concept of marginal efficiency.

Keynes also notes that increased investment in particular capital assets increases supply prices and reduces demand prices, causing a decline in marginal efficiencies; an increase in output thus leads to investment in assets with lower rates of return. At some point the marginal efficiency of money will make investment in money as profitable as the purchase of capital assets. At this point the rate of interest equals the marginal efficiency of capital, and any further increase in output would confirm Keynes's 'general principle' that any further expansion in output gluts the market, for increased income is not spent but held in the form of money which becomes a 'generalised sink for purchasing power'.

The question that distinguishes Keynes's theory is thus why money's liquidity premium does not fall as output expands, for this is what prevents investment from rising by just the amount to fill the gap created by the propensity to consume being less than one. To describe these 'essential properties of interest and money', Keynes departs from Mill's position that money is just another commodity. When money is the debt of the banking system its price and quantity behaviour will differ from physical commodities, for it has no real costs of production nor real substitutes. Thus an asset which has a negligible elasticity of production and substitution with respect to a change in effective demand, will have a rate of return which responds less rapidly to an expansion in demand. As long as the rate of interest falls less rapidly than the marginal efficiencies of capital assets, its rate will be the one which sets the point at which further expansion creates losses.

Thus the propensity to consume shows that investment will have to increase by the amount of the gap between incomes and expenditures as incomes rise if entrepreneurs are not to make losses, while the marginal efficiency of capital and liquidity preference in a monetary production economy explain why the behaviour of the rate of interest relative to the marginal efficiency of capital makes it unlikely that the rate of investment should adjust by just that amount. Since entrepreneurs maximize monetary returns, not employment or physical output, there is no reason why their investment decisions should lead to an equilibrium at full employment. Keynes's explanation

of the limit to the level of employment permits any level as a stable equilibrium, including full employment; it is thus more general than the classical Say's Law position, in which the only stable equilibrium was the limit set by full employment as given in the labour market.

See Also

► Say's Law

Bibliography

Keynes, J.M. 1930. *A treatise on money.* Reprinted in Keynes (1971).
Keynes, J.M. 1934a. Letter to R.F. Kahn, 13 April. Reprinted in Keynes (1973b).
Keynes, J.M. 1934b. Letter to P. Sraffa, undated. Reprinted in Keynes (1979).
Keynes, J.M. 1934c. *Poverty in plenty: Is the economic system self-adjusting?* Reprinted in Keynes (1973b).
Keynes, J.M. 1936a. *The general theory of employment, interest and money.* Reprinted in Keynes (1973a).
Keynes, J.M. 1936b. Letter to A. Lerner, 16 June. Reprinted in Keynes (1979).
Keynes, J.M. 1936c. Letter to R.F. Harrod, 30 August. Reprinted in Keynes (1973c).
Keynes, J.M. 1937a. *The theory of the rate of interest*. Reprinted in Keynes (1973c).
Keynes, J.M. 1937b. *Alternative theories of the rate of interest*. Reprinted in Keynes (1973c).
Keynes, J.M. 1971–83. *The collected writings of John Maynard Keynes,* ed. D. Moggridge. London: Macmillan for the Royal Economic Society: 1971. Vols. 5 and 6. *A treatise on money* (1930). 1973a. Vol. 7. *The general theory of employment, interest and money* (1936). 1973b. Vol. 13. *The general theory and after: Part I – preparation.* 1973c. Vol. 14. *The general theory and after: Part II – defence and development.* 1979. Vol. 29. *The general theory and after – a supplement.*
Malthus, T.M. 1821. Letter from Malthus to Ricardo, 7 July. Reprinted in Ricardo (1952), 9–10.
Mill, J.S. 1874. On the influence of consumption on production. *Essays on some unsettled questions of political economy*, 2nd ed, reprinted Clifton, ed J.S. Mill. Clifton: A.M. Kelley, 1974.
Ricardo, D. 1952. *Works and correspondence of David Ricardo*, vol. 9, ed. P. Sraffa with the collaboration of M. Dobb. Cambridge: Cambridge University Press.
Say, J.B. 1855. *A treatise on political economy,* 6 American ed. Philadelphia: J.B. Lippincott.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford University Press, 1976.

Effective Protection

W. M. Corden

The effective rate of protection is the rate of protection provided to the value added in the production of a product. Let the effective price be defined as the domestic price of a unit of value added. Then the effective rate of protection (henceforth, ERP) is the proportional increase in the effective price made possible by tariffs and other measures. It is to be contrasted with the nominal tariff and (more generally) nominal rate of protection, which refers to the proportional increase in the nominal price. If the only policy instruments are tariffs, the ERP depends not only on the nominal tariff on the commodity concerned but also on the tariffs on the inputs and on the input coefficients.

Consider the simple case of an importable product, j, which has only a single input, also an importable, i. There are no taxes and subsidies affecting j and i other than the import tariffs. The formula for the ERP for the activity producing j is then

$$g_j = \frac{t_j - a_{ij}t_i}{1 - a_{ij}}$$

where gj is the ERP, tj is the tariff on j, ti is the tariff on i, and aij is the share of i in the cost of j in the absence of tariffs.

This shows that if tj = ti, then gj = tj. It is common for input tariffs to be low relative to final goods tariffs, that is, tj > ti, and in that case gj > tj, an important result, since it shows that effective rates tend to be higher than nominal rates. A rise in the input tariff clearly reduces effective protection for the using industry, even though it raises protection for the input-producing activity.

Actual measurements involve using a'_{ij}, which is the input share that results after the tariffs have raised both the domestic final good price and the domestic input price. The connection between the input share before tariffs are imposed (aij) and after (a'_{ij}) is as follows:

$$a'_{ij} = a_{ij}\frac{1+t_i}{1+t_j}$$

and from this, and the formula for the ERP given above, one can obtain the formula which is commonly used in empirical studies, namely:

$$g_j = \frac{1-a_{ij}}{\frac{1}{1+t_j} - \frac{a'_{ij}}{1+t_i}} - 1.$$

The effective protection concept can be extended to allow for all taxes and subsidies affecting tradeable goods, i.e. all importables and exportables. An export subsidy raises the domestic price of an exportable; if the input is an exportable, then ti represents the rate of subsidy, and if the activity for which the ERP is being calculated is an exportable, then tj can represent the subsidy. Similarly, export taxes, production taxes, consumption taxes, and production and consumption subsidies can be allowed for. A production tax or subsidy for the final product will affect tj, while a consumption tax or subsidy for the input will affect ti. Thus the ERP measure allows a single figure to sum up the net result of various trade and non-trade taxes and subsidies affecting any particular activity.

The ERP measurements revealed at an early stage the high protection that developed countries provided for final processing of primary products even in cases where nominal tariffs were low, the reason being the duty-free entry of the basic materials. The measurements also bring out the negative effective protection provided for exports in many countries: the exports receive no subsidies or other assistance, i.e. tj = 0 for most exports, but import tariffs on their inputs make their tis positive. It was also noted that tariff reductions are not always what they seem: an offer of tariff cuts at an international negotiation may actually raise the ERP for some domestic industries.

Much attention has been given in the literature to the discovery of negative value added, a discovery which was a by-product of effective protection calculations. There are cases where the free trade price of the final product is less than the free trade price of its inputs, so that under free trade the effective price would be negative. One possible reason for this phenomenon is that transport costs on inputs may be much higher than those on the final product. There would then be no production of the final good under free trade. But a sufficiently high tariff on the final good relative to the tariff on the input could make the effective price domestically positive, so that domestic production begins. The rate of protection is then infinite, and algebraically the calculation of the ERP will yield a negative figure. Clearly domestic production of a product where the cost of imported inputs exceeds the free trade price of the final product is an extreme form of waste.

General Equilibrium

The next step is to put ERPs into a general equilibrium framework. One can imagine a scale of effective rates, which will include ERPs for all traded activities, including both exportables and importables. The scale will give some indication of the direction in which resources have been pulled by the protective structure. Of course, actual resource movements will also depend on production substitution elasticities, that is, on the whole general equilibrium system, so that the scale is only indicative of resources movement effects. The crucial point is that, in general equilibrium, relative ERPs matter, not absolute rates. This is simplest to see in a model with only two activities where both may be obtaining positive ERPs (if one is an export, this implies it is getting an export subsidy), and resources will then tend to move into the activity with the relatively higher ERP.

There are complications, and it has been shown in the literature that one can produce paradoxes. For example, in a three-activity model, with A and B complementary in a general equilibrium sense, protection of A may expand B even though B may get a lower ERP than C. It must also be remembered that relative nominal rates will determine the direction in which consumption is pulled or distorted.

The idea of the scale of ERPs was first presented in Corden (1966), where it was said that

Assuming normal non-zero substitution elasticities in production, [the scale of effective rates] tells us the direction in which this structure causes resources to be pulled as between activities producing traded goods. Domestic production will shift from low to high effective-protective-rate activities.

This was too strong. It is true in a rather special model, later set out formally in Jones (1975), but more generally, there are various 'paradoxical' circumstances where it need not be true. Several articles have explored such possibilities, and examples are expounded in Corden (1971).

At the general equilibrium level the important and somewhat complex issue also arises of whether particular traded goods activities are protected relative to non-tradeables, and what the role of the exchange rate is. The basic point is that the imposition of a protective structure which is generally positive will tend to draw resources out of non-tradeables, and if the nominal rates are also mainly positive, divert consumption towards non-tradeables. Assuming balance initially, protection then results in excess demand for non-tradeables. Balance would be restored by a rise in the price of non-tradeables, relative to the free trade prices of tradeables, this being a real appreciation. It could be brought about with a fixed nominal exchange rate combined with an absolute rise in the price of non-tradeables, or by a nominal appreciation when the price of non-tradeables is constant, possibly because the nominal wage is given.

The usual expositions assumed the average price-level of non-tradeables constant, and stressed the exchange rate adjustment that then needs to be associated with a change in protection levels. When this adjustment is taken into account one can obtain a net protective rate which shows whether a particular activity is protected relative to non-tradeables. For example, a particular activity may obtain an ERP of 10 per cent but, if the protection for it and all other activities were removed, there might have to be a devaluation which is equivalent to a uniform tariff and export subsidy of, say, 15 per cent. In that case this activity would have obtained a higher effective price under free trade, so that the system of protection has provided it with negative net effective protection. Relative to non-tradeables, it has been anti-protected. If its resources were primarily mobile into and out of non-tradeables, it would expand as a result of a movement towards free trade.

Key Assumptions

The theory of effective protection, at least in its more formal version, makes a number of assumptions. The first is the small country assumption, namely the assumption that the country concerned faces given prices of its exports and imports (the terms of trade being exogenous). The second assumption is that for all tradeable goods some trade remains, so that domestic prices are determined by the given world prices as modified by tariffs, export subsidies and other interventions. Thirdly, imports are assumed to be perfect substitutes for the import-competing goods for which the ERPs are calculated. Finally, it is assumed that there are fixed coefficients between final outputs and traded inputs, even though substitution between the domestic factors that contribute to value added can be allowed.

Much theoretical work has gone into exploring the implications of removing the last assumption, though all the others are also important. It does not follow that calculations of ERPs are meaningless when these assumptions do not hold, but figures must be interpreted with care, as discussed in Corden (1971).

Normative Implications

Do ERPs have normative significance? The formal theory of effective protection and tariff structure was developed with the focus on a question of positive economics: namely, how does a protective structure affect the allocation of resources? But the great interest in the theory and the widespread activity in making calculations has been motivated by a concern with normative issues. It must be stressed again that only relative effective rates matter. Knowing a single effective rate on its own sheds no light on either positive or normative

implications. The frequent assumption, often only implicit, has been that free trade with appropriate exchange rate adjustment would be the optimal situation, and that non-uniform effective rates therefore impose a production cost of protection – that is, a welfare loss through a distortion in resource use. Large divergences are then an indication of a high cost of protection. Furthermore, the structure of effective protection gives then a guide to the welfare (or efficiency) effects of tariff changes: a change that reduces a divergence between effective rates is likely to reduce the cost of protection.

A practical implication is that if there is to be gradual tariff reduction without extra costs being imposed during the process, any increase of such divergences should be avoided. This will be so, for example, if high effective rates are always reduced first. In a three-product model, with industry A getting 0 per cent, industry B 20 per cent and industry C 50 per cent, a reduction in B's effective rate first would increase the divergences between the ERPs on B and C, so that the ERP of C should be reduced first. This is the concertina method of tariff reduction, but may have quite complicated implications in terms of nominal tariffs. Radial (uniform across-the-board) reductions would also avoid divergences being increased. Finally, it must be remembered that nominal tariffs affect the pattern of consumption whether by final users, or in use of inputs, so that divergences in nominal tariffs determine the consumption cost of protection.

If there are other (non-trade) distortions in the economy, a tariff distortion may actually be offsetting. Thus, if an industry is established on the basis of a very high tariff (relative to other industries), so that a positive cost of protection might be expected, there may be a gain if (for example) the industry uses labour for which it has to pay a wage that exceeds its opportunity cost owing to distortions in the labour market. When such non-trade distortions are prevalent one cannot use effective rates on their own as indicators of which activities should expand and which decline if resource allocation is to improve. The broader concept of domestic resource cost has been developed to take all distortions into account.

Practical Problems

The calculation of ERPs and their use as a guide to policy has become very widespread, especially in developing countries. But all sorts of practical problems arise in the calculations, essentially because the assumptions of the formal theory do not hold, and many ways have been devised to deal with these problems. The problems can only be listed here, but they are important for practitioners. For more details, see Corden (1975) and Balassa et al. (1982).

When quotas are the principal method of protection, comparisons between domestic and world market prices must be made in order to obtain the implicit nominal rates of protection which must be the starting point for any calculations. When tariffs alone are relevant there may be tariff redundancy, so that, again, price comparisons must be made; a difficulty here is that the quality of the local product and the import may differ. Available input-output coefficients in most countries are rarely sufficiently disaggregated for the ERP calculations. There is a need for tariff averaging, and this has built-in biases.

A decision has to be made as to how to treat non-traded inputs into the tradeable products for which ERPs are calculated. This last issue has given rise to much theoretical discussion (on which see Corden 1971, 1975). The correct method appears to be very complicated: lump the non-traded and primary-factor content of non-traded inputs with value added, but group the traded-input content of non-traded inputs with traded inputs. Tariffs on traded inputs into non-traded inputs then reduce the ERP for the final product.

The Substitution Problem

By far the most sophisticated theoretical work has gone into the 'substitution problem'. This involves removing the assumption of fixed coefficients between the final output and the produced traded inputs. Thus substitution between traded inputs and the primary factor content of value added is allowed for. Two distinct issues then arise.

First, suppose that the production functions are separable, so that substitution between traded inputs and the various primary factors, for example, labour and capital, is 'unbiased'. In that case, the concept of value added retains a clear meaning. One can think of a 'value added product' which is combined with traded inputs in varying proportions (depending on the input tariff and the final good tariff, among other things), to make a final product. Since the ERP is the proportional increase in the effective price, which is the price of this 'value added product'. ERP also then has a clear meaning. But the problem remains that measurements based on the coefficients after tariffs have been imposed (which is what the data yield) will have a bias, reflecting the substitution effects. It can be shown that the tendency will always be to overstate the 'true' ERPs. The problem is then one of inevitable measurement error. Since one is interested in the relative position in the scale of effective rates and in the divergences between ERPs, it is relevant that the measurement error will differ between ERPs, depending on production functions and relationships between final goods and input tariffs. Fortunately, there is some possibility that this complication may not be important in practice.

The second issue is more fundamental. If production functions are not separable, so that substitution is 'biased', the whole concept of the 'value added product' and hence of ERP is thrown into doubt. The question is whether 'value added' has a meaning. One really needs to assume that, on a probability basis, the bias is generally zero.

Are General Equilibrium Models Preferable?

Another basic criticism of the ERP concept and of all the resources that have gone into calculations of ERPs can be made. It has been pointed out that a scale of ERPs is an imperfect and possibly misleading indicator of resource allocation movements. Actual resource pulls also depend on supply elasticities, on production functions, and on a whole lot of complex interactions which have been analysed in the literature, but which deprive the scale of effective rates of any simple significance. Various paradoxes have been shown to be possible – for example, that resources will be drawn into a low ERP activity out of a high ERP one under particular factor-intensity and relative tariff conditions. The conclusion of some critics has been either that no measurements are any use or that one might as well use only nominal rates. Another view is that the best approach is to use computable general equilibrium models, and these make ERPs redundant.

The answer must be that if the data and estimates for complete general equilibrium models are available – and sufficiently disaggregated with respect to activities or industries to be policy-relevant – there is indeed no need to calculate ERPs. The latter contain some information, taking into account input tariffs, and so on, but pause half-way to the complete answer. The case for ERPs calculations and their use for policy must be that the data and estimated functional relationships required for complete and detailed general equilibrium models do not usually exist, and certainly not in sufficiently disaggregated form, so that ERPs, which are feasible to calculate, give some indication of possible resource pulls and costs of protection. The extensive theoretical work is designed to indicate the direction of various probable biases and to bring out the stringent assumptions required for firm conclusions to be reached from the data.

The Literature

Until the mid–1960s the vertical relationships between tariff rates derived from the input-output relationships between products were completely neglected in the literature of trade theory. In fact tariff theory was either narrowly partial equilibrium, focusing on just one vertically integrated product, or consisted of two-sector general equilibrium models. A major feature of the theory of tariff structure was not just to bring out the relevance of input tariffs but also to focus on the horizontal general equilibrium relationships when there are more than two products.

With regard to the ERP concept itself, while there were early precursors, the first extended exposition was in Barber (1955), the first systematic theoretical papers were a 1965 paper of Johnson's, reprinted in Johnson (1971), and Corden (1966). The latter paper opened up various general equilibrium issues, the significance of the scale of effective rates, the problem of non-traded inputs, the substitution problem, and so on, and later a systematic and more complete exposition was presented in Corden (1971), which also contains a history of the ERP concept and references to various precursors. Pioneering empirical work was done in Balassa (1965) and Basevi (1966). Later Balassa became a sponsor of major multi-country empirical studies (Balassa et al. 1971, 1982), and these volumes also contain extensive reviews by Balassa of theoretical and measurement issues.

The central theoretical issues of the meaning of ERPs have been discussed in numerous papers subsequent to the early work. Particularly to be noted are Jones (1975) and Ethier (1977). In addition, there have been several articles on the 'substitution problem', beginning with Jones (1971), a paper reprinted in Corden (1971), and Ethier (1972), followed by papers by Bruno, by Khang and by Bhagwati and Srinivasan, all in the Journal of International Economics of 1973.

See Also

- ▶ Free Trade and Protection
- ▶ International Trade
- ▶ Quotas and Tariffs

Bibliography

Balassa, B. 1965. Tariff protection in industrial countries: An evaluation. *Journal of Political Economy* 73(6): 573–594.

Balassa, B., et al. 1971. *The structure of protection in developing countries*. Baltimore: Johns Hopkins Press.

———. 1982. *Development strategies in semi-industrial countries*. Baltimore: Johns Hopkins Press.

Barber, C.L. 1955. Canadian tariff policy. *Canadian Journal of Economics and Political Science* 21(November): 513–530.

Basevi, G. 1966. The US tariff structure: Estimation of effective rates of protection of US industries and industrial labor. *Review of Economics and Statistics* 48: 147–160.

Bruno, M. 1973. Protection and tariff change under general equilibrium. *Journal of International Economics* 3(3): 205–225.

Bruno, M., C. Khang, A. Ray, J.N. Bhagwati, and T. Srinivasan. 1973. The theory of effective protection in general equilibrium: A symposium. *Journal of International Economics* 3(3): 205–281.

Corden, W.M. 1966. The structure of a traiff system and the effective protective rate. *Journal of Political Economy* 74: 221–237.

———. 1971. *The theory of protection*. Oxford: Oxford University Press.

———. 1975. The costs and consequences of protection: A survey of empiricial work. In *International trade and finance: Frontiers for research*, ed. P.B. Kenen. Cambridge: Cambridge University Press.

Ethier, W.J. 1972. Input substitution and the concept of the effective rate of protection. *Journal of Political Economy* 80(1): 34–47.

———. 1977. The theory of effective protection in general equilibrium: Effective-rate analogues of nominal rates. *Canadian Journal of Economics* 10(2): 233–245.

Grubel, H.G., and H.G. Johnson (ed). 1971. *Effective tariff protection*. Geneva: Graduate Institute of International Studies.

Johnson, H.G. 1971. *Aspects of the theory of tariffs*. London: George Allen & Unwin.

Jones, R.W. 1971. Substitution and effective protection. *Journal of International Economics* 1(1): 59–81.

———. 1975. Income distribution and effective protection in a multicommodity trade model. *Journal of Economic Theory* 11(1): 1–15.

'Effectual Demand' in Adam Smith

Carlo Panico

Smith's notion of 'effectual demand' is still the subject of several discussions dealing with the role of demand in classical and neoclassical theories of price and distribution and with the influence of demand on 'division of labour' and economic progress. Smith defined 'effectual demand' as the 'demand of those who are willing to pay the natural price of the commodity, or the whole value of rent, labour and profit, which must be paid in order to bring it thither' (Smith 1776,

vol. 1, p. 58). According to him, when the quantity of any commodity brought to market falls short of the effectual demand, those who demand it.

Cannot be supplied with the quantity they want. Rather than want it altogether, some of them will be willing to give more. A competition will immediately begin among them, and the market price will rise more or less above the natural price (ibid.).

On the other hand, 'when the quantity brought to market exceeds the effectual demand,... the market price will sink more or less below the natural price' (p. 59), whereas 'when the quantity brought to market is just sufficient to supply the effectual demand and no more, the market price naturally comes to be ... the same with the natural price' (ibid.).

'Effectual demand' is thus defined as the demand for any *individual* commodity, corresponding to the natural price for it. It was a *long-period* concept, since it was associated with those prices which allow the payment of wages, rents and profits at their natural levels, and which hold when in all industries productive capacity is fully adjusted and a uniform rate of profits is earned (see Smith 1776, vol. 1, pp. 59–65).

The definition of 'effectual demand' was introduced in dealing with the adjustment process between demand and supply. This process was conceived to occur on a *single* market assuming as known the natural prices of that and all other commodities. The process of adjustment implies, therefore, a *prior* determination of distributive variables and of all natural prices, *associated with given levels of effectual demand in each industry.* Smith's notion of 'effectual demand' thus refers as much to a specific industry as to the whole economy: it can be seen as a 'micro' and a 'macroeconomic' concept.

The study of effectual demand involves a description of how the working of competition enforces natural prices but does not constitute a theory of what determines them. Smith never derived demand-functions for any commodity. 'Effectual demand' represented a point, and no attempt was made to determine the magnitude of the rise (fall) in demand when the price falls below (rises above) its natural level. He thus used a different notion from that implied by demand-curves in neoclassical theory, which requires a specific ordering between *each* price-quantity point. ... The theory does not regard these points as results of accidental and temporary deviations of the quantity supplied from the 'normal' level, but rather as determinate points likely to emerge from a repetition of events (Garegnani 1983, p. 310).

Smith's notion of effectual demand has been recalled by those who, following Sraffa's rehabilitation of the surplus approach of the classical political economists (Sraffa 1960), have proposed to separate the analysis of price and distribution from that of the levels of output and demand. Within this approach, *given the level and the composition of output* and one distributive variable, it is possible to determine the 'socially necessary' technique, the other distributive variables and natural prices. The levels of output and demand in each industry, taken as given, represent long-period values, since they are associated with fully adjusted productive capacity and uniform rates of profit in all industries.

The analysis of the classical tradition is characterized by integration between historical, institutional and economic factors. This approach is applied to the analysis of the level and composition of demand. The analysis of the aggregate level is related to Say's law, whose acceptance is an open option in classical political economy. Among the elements affecting the composition of demand, two groups of factors appear to emerge in Smith's writings. First of all, *objective* factors, like the degree of development of the economy and the distribution of income among different classes of society. Secondly, *subjective* factors, which are influenced by customs, social rules and fashion. The limited attention paid to substitution within the bundle of commodities demanded by different income groups suggests a minor role attributed to this factor, without denying the possibility of its further analysis, carried out case by case.

Marx analysed the factors influencing demand in a similar way. His stress was on objective factors, that is on the ratio of total surplus-value to wages and the proportions in which the surplus-

value is split up among profits, interests, ground rents, taxes, etc. (see Marx 1894, pp. 181–2). Given the historically achieved degree of development of the economy (whose analysis is not based on the acceptance of Say's law) and the distribution of income, it is possible to determine the average level of demand for different commodities from each class or social group. The total consumption expenditure of each class is an increasing function of the income earned (Marx 1894, pp. 188–9), while the composition of its consumption is influenced by habits and rules which, over a certain historical period, are dominant within that class. Limited possibilities of substitution within the bundle of commodities demanded by each class are recognized, and again appear left to be studied case by case.

The working class must find at least the same quantity of necessities on hand if it is to continue living in its accustomed average way, although they may be more or less differently distributed among the different kinds of commodities . . . The same, with more or less modification, applies to other classes (Marx 1972a, pp. 188).

Besides, Marx pointed out that the analysis of demand has to recognize the distinction between the part coming from consumers and that coming from entrepreneurs requiring means of production in order to meet what he called the need for commodities in the market, depending on the 'actual social needs of the different classes and on the income available to them' (Marx 1894, pp. 188–9).

Some remarkable similarities can be found between this approach and that followed by Keynes in the *General Theory*. In chapters 8 and 9 of this work, the factors affecting aggregate consumption are examined in an analysis which is separate from that determining prices and distribution, and which pays hardly any attention to substitution within the bundle of commodities demanded for consumption, a factor to which a secondary role appears to be attributed. According to Keynes (see 1936, pp. 90–95), total consumption depends partly on total income, partly on other objective circumstances, like the interest rate, and partly on subjective factors, which 'include those psychological characteristics of human nature and those social practices and institutions' (p. 91), which are unlikely to change over limited periods of time except in abnormal or revolutionary circumstances, and which it is necessary to consider 'in an historical inquiry or in comparing one social system with another of a different type' (ibid.). Talking of the interest rate, Keynes concluded that its influence on consumption is open to a great deal of doubt. . . .[Its influence] is complex and uncertain, being dependent on conflicting tendencies . . . Substantial changes in the rate of interest tend to modify social habits considerably, thus affecting the subjective propensity to spend – though in which direction it would be hard to say, except in the light of actual experience (p. 93).

Thus, as in classical tradition, the actual influence of the factors considered is evaluated by Keynes according to the historical circumstances considered, taking into account that their influence may be uncertain in its intensity and direction. The integration between economic and institutional and social factors also emerges in the analysis of the influence of subjective factors (pp. 107–112), whose relative strength will vary enormously according to the institutions and organisation of the economic society which we presume, according to habits formed by race, education, convention, religion and current moral, according to present hopes and past experience, according to the scale and technique of capital equipment, and according to the prevailing distribution of wealth and the established standards of life (p. 109).

However, the principle of substitution and that of diminishing marginal returns play a primary role in Keynes's analysis of investment in the *General Theory*. In this respect, Keynes said, 'I am simply accepting the usual theory of the subject' (Keynes 1973, p. 615), 'meaning exactly the same as Marshall . . . means' (p. 630). Yet, alongside this neoclassical element, Keynes referred to other factors influencing investment, like the present and expected level of effective demand (see Keynes 1936, p. 147), which may come from the private or the public sector and may affect what he

called 'the state of long-term expectation'. The analysis of investment of the *General Theory* may thus suggest some elements to develop a theory of demand within classical tradition.

One element is that 'the state of long-term expectation is often steady' (Keynes 1936, p. 162), since factors like the institutional environment and government policies do not only influence it, but also 'exert their compensating effects' on its fluctuations, together with factors related to the maintenance of the efficiency of capital goods. Within this line, government policies, and industrial policy in particular, relations between industry and finance, industrial relations and the history of competitiveness and technological changes are to be seen as relevant factors affecting the prevailing state of long-term expectation (see Eatwell 1983, p. 283).

Another element is that there may be 'short-period changes in the state of long-term expectation' (Keynes 1936, p. 164) due, among other things, to reactions of investors during the transition process from one state of long-term expectation, 'which has its definite corresponding level of long-period employment' (ibid., p. 48) with fully adjusted capacity, to another to which a new long-period position corresponds. This process was described by Keynes in chapter 5 of the *General Theory* (1936, pp. 46–50), where he concluded that 'a mere change in expectation is capable of producing an oscillation of the same kind of a shape as a cyclical movement, in the course of working itself out' (p. 49). This chapter points out the possibility of presenting a long-period analysis of demand and output, which is integrated with an analysis of the cyclical movements of the economy.

Smith's notion of 'effectual demand' thus appears a fruitful concept linking the classical theory of prices and distribution and that of output and demand. The historical elements present in the latter theory underline an outstanding feature of Smith's and of classical political economists' work, i.e. that the analysis of output and demand is part of the analysis of concrete 'historical processes' of accumulation which, as said above, can show cyclical fluctuations around the main trend.

Bibliography

Eatwell, J.L. 1983. The long-period theory of employment. *Cambridge Journal of Economics* 7: 269–285.

Garegnani, P. 1983. The classical theory of wages and the role of demand schedules in the determination of relative prices. *American Economic Review: Papers and Proceedings* 73: 309–313.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Keynes, J.M. 1973. *The general theory and after*. Part I: *Preparation*. Vol. XIII of *The collected writings of J.M. Keynes*, ed. D. Moggridge. London: Macmillan.

Marx, K. 1894. *Capital*, vol. 3. London: Lawrence & Wishart, 1972.

Marx, K. 1910. *Theories of surplus value*, vol. 3. London: Lawrence & Wishart, 1972.

Smith, A. 1776. *An inquiry into the nature and the causes of the wealth of nations*, 2 vols, ed. E. Cannan. London: Methuen, 1930.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Efficiency Bounds

Han Hong

Abstract

In large sample analysis, the performances of estimators can be approximated by the asymptotic variances. In parametric models, maximum likelihood estimators often achieve the efficient Cramer–Rao lower bound, while efficient GMM estimation can be achieved by choosing the weighting matrix and the instruments optimally. Semiparametric efficiency bound is defined by the supremum of the Cramer–Rao bounds for all parametric models that satisfy the semiparametric restrictions. The efficiency bounds for asymptotically linear semiparametric estimators are given by the variances of the efficient influence functions, which are the projections of the linear influence functions onto the tangent spaces of the semiparametric models.

Keywords

Asymptotic efficiency; Asymptotic variance; Asymptotically linear estimators; Average risk optimality; Cramer–Rao lower bound; Delta method; Efficiency bounds; Euler equations; Generalized method of moment; Hodges' estimator; Invariance principle; Maximum likelihood; Mean square errors; Measurement error models; Measures of; Closeness; Minimax optimality; Parametric models; Partial linear model; Semiparametric efficiency bound; Semiparametric models; Unbiased estimators

JEL Classifications

C14

Oftentimes we want to compare estimators. For a given parameter in which we are interested, there are typically many estimators that can estimate it consistently. We need to choose the best estimator, or the estimator that is the closest to the true parameter value. The mean square error (MSE), $E\left(\hat{\theta}-\theta\right)^2$, is frequently used as a measure of closeness. However, there can be many other various measures of closeness, and often they do not agree with each other. See, for example, Amemiya (1994, pp. 116–24).

Even with a given measurement of closeness, such as the MSE, it is typically not possible to rank two estimators. For two estimators X and Y of θ, X is better than Y only if $E(X-\theta)^2 \leq E(Y-\theta)^2$ for all $\theta \in \Theta$. An estimator that is not dominated by another estimator in the above sense is called admissible.

A uniformly 'most' efficient estimator does not exist. To find an efficient estimator, one needs to confine the analysis to a limited class of estimators, such as unbiased estimators or equivariant estimators. Alternatively, one can rely on a subjective strategy such as average risk optimality which requires a prior distribution over the parameter space, or use a pessimistic and risk-averse approach such as minimax optimality.

In large sample analysis, the performance measures of estimators can often be approximated by their asymptotic distribution. Under suitable regularity conditions, many estimators are consistent and converge to the true parameter values at $\sqrt{n}$ rate. These estimators can be compared based on their asymptotic variance. The notation of efficiency bound usually refers to the largest lower bound for the variances that can be achieved by $\sqrt{n}$ consistent and asymptotically normal estimators under suitable regularity conditions.

Asymptotic Efficiency in Parametric Models

In parametric models, the variance of an unbiased estimator has to be larger than the Cramer–Rao lower bound, which is defined as the inverse of the information matrix:

$$V\left(\hat{\theta}\right) \geq -\left(E\frac{\partial^2 \log L}{\partial \theta^2}\right)^{-1},$$

where L is the likelihood function. Proofs of this result can be found, for example, in Amemiya (1994, pp. 138–39; 1985, pp. 14–17). A consistent estimator is said to be asymptotically efficient if its asymptotic variance achieves the Cramer–Lao lower bound. Under suitable regularity assumptions such as those given in Theorem 4.1.3 in Amemiya (1985), the maximum likelihood estimator is asymptotically efficient.

There exist super-efficient estimators whose asymptotic variances are smaller than the Cramer–Rao lower bound on a set of parameter θ with Lebesgue measure zero, such as Hodges's estimator defined as

$$w_T = \begin{cases} 0 & \text{if} \quad \left|\hat{\theta}\right| < T^{1/4} \\ \hat{\theta} & \text{if} \quad \left|\hat{\theta}\right| \geq T^{-1/4} \end{cases}$$

where $\sqrt{T}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N(0, v(\theta))$. One can show that $\sqrt{T}(w_T-\theta) \xrightarrow{d} N(0, v(\theta))$ if $\theta \neq 0$ and $\sqrt{T}(w_T) \xrightarrow{d} 0$ if $\theta = 0$. However, the better behaviour of w_T at $\theta = 0$ comes at the expense of erratic

behaviour when θ is close to 0. See for example, van der Vaart (1999, p. 110).

A common alternative to maximum likelihood is generalized method of moment estimators (GMM). Its asymptotic efficiency is extensively discussed in Newey and McFadden (1994). While GMM estimators are less efficient than maximum likelihood (see, for example, the proof in Newey and McFadden (1994, p. 2163), oftentimes they are easy to compute, especially when maximum likelihood is computationally infeasible. For a given set of unconditional moment conditions, a proper choice of the weighting matrix or the linear combination matrix minimizes the asymptotic variance. For a given set of conditional moment conditions, a proper choice of instruments can also minimize the asymptotic variance.

A GMM estimator can be formed from the over-identified moment conditions $Em(z;\theta) \equiv 0$ by minimizing a quadratic form based on a weighting matrix W:

$$\frac{1}{T}\sum_{t=1}^{T} m\left(z_t;\hat{\theta}\right) W \frac{1}{T}\sum_{t=1}^{T} m\left(z_t;\hat{\theta}\right).$$

The resulting estimator has asymptotic variance $(G\prime WG)^{-1}(G'W\Omega WG)(G\prime WG)^{-1}$, where $G = E\frac{\partial}{\partial\theta}m(z;\theta)$ and $\Omega = Var(m(z;\theta))$ Hansen (1982) showed that the optimal choice of $W = \Omega^{-1}$, which equates $G'WG = G'W\Omega WG$. In this case the asymptotic variance is reduced to $(G'\Omega^{-1}G)^{-1}$.

Alternatively, a set of over-identified moment conditions $Em(z;\theta) \equiv 0$ can be translated into a set of exactly identified moment conditions by a linear combination matrix $AEm(z;\theta) \equiv 0$. Given A, the resulting method of moment estimator that equates $A\sum_{t=1}^{T} m\left(z_t;\hat{\theta}\right)$ to zero has asymptotic variance $(AG)^{-1}(A\Omega A')(G'A')^{-1}$. As a rule of thumb, the optimal choice of A should simplify this asymptotic variance, by equating $AG = A\Omega A' = G'A'$. The resulting optimal $A = G'\Omega^{-1}$ gives rise to the same asymptotic distribution as the above optimally weighted GMM estimator of Hansen (1982), which minimizes

$$\frac{1}{T}\sum_{t=1}^{T} m\left(z_t;\hat{\theta}\right) \Omega^{-1} \frac{1}{T}\sum_{t=1}^{T} m\left(z_t;\hat{\theta}\right).$$

Many economic models, such as those based on Euler equations, are stated in terms of conditional moment conditions of the form $E(m(z;\beta)|\, x) = 0$ for almost all x. These conditional moment conditions can be translated into exactly identified unconditional moment conditions using an instrument matrix $A(x) : EA(x)m(z;\beta) = 0$. The question arises as to what is the optimal instrument matrix $A(x)$. For a given choice of $A(x)$, the resulting method of moment estimator that equates $\frac{1}{T}\sum_{t=1}^{T}(x_t)m(z_t;\beta) = 0$ has asymptotic variance $(EA(x)G(x))^{-1}EA(x)\Omega(x)A(x)'\ (EG(x)'\ A(x)')^{-1}$, where $G(x) = E\left(\frac{\partial}{\partial\theta}m(z;\theta)|\, x\right)$ and $\Omega(x) = Var(m(z;\beta)|\, x)$. We can then equate

$$EA(x)G(x) = EA(x)\Omega(x)A(x)'$$

to obtain the optimal instrument matrix $A(x) = G(x)'\Omega(x)^{-1}$. The resulting efficient asymptotic variance is therefore $(EG(x)'\Omega(x)^{-1}G(x))^{-1}$.

Formal proofs of these derivations can be found in, for example, Newey and McFadden (1994). Estimators that achieve these efficiency bounds typically involve two-step or multi-step procedures and possibly nonparametric methods, such as Newey and Powell (1990).

Asymptotic Efficiency in Semiparametric Models

Semiparametric models are extensions of parametric models where some components are specified nonparametrically with unknown functional forms. Generalized method of moment models are semiparametric models if the data-generating process is not fully specified. A partial linear model is another example. Other popular semiparametric models are surveyed in Powell (1994).

Intuitively, the variance of an estimator for a semiparametric model should be larger than the Cramer–Rao lower bound for any parametric sub-model that satisfies the semiparametric restrictions. The semiparametric efficiency bound is

therefore defined to be the supremum of the Cramer–Rao bounds for all parametric models that satisfy the semiparametric restrictions. Extensive results for semiparametric efficiency bounds are developed in, among others, Bickel et al. (1993) and Newey (1990). In this section we give a brief summary of some of the results presented in Newey (1990). The next section will apply these results to a particular estimation problem.

Because of pathological cases such as the super-efficient estimator, the semiparametric efficiency bound is used to provide a lower bound only for *regular* estimators. Consider a parameter of interest that is a smooth function of the underlying parametric path: $\beta(\theta)$. A regular estimator $\hat{\beta}$ is one where for each θ_0 the limiting distribution of $\sqrt{T}\left(\hat{\beta} - \beta(\theta_T)\right)$ does not depend on θ_T as long as $\sqrt{T}(\theta_T - \theta_0)$ is bounded. The super-efficient estimator is not regular.

Most estimators in econometrics are asymptotically linear, in the sense that they have an influence function representation as

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi(z_t) + o_p(1).$$

In particular, almost all econometric estimators asymptotically solve some moment conditions $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m\left(z_t; \hat{\beta}\right) = o_p(1)$, in which case the linear influence function is given by $\psi(z_t) = -G^{-1}m(z_t; \beta)$ for $G = E\frac{\partial}{\partial \beta} m(z_t; \beta)$.

Asymptotically linear estimators are regular if and only if for all parametric sub-models $\frac{\partial}{\partial \theta}\beta(\theta) = E\psi S_\theta'$. When $\psi(z_t) = -G^{-1}m(z_t; \beta)$, this follows from differentiating $E_\theta m(z; \beta(\theta)) = 0$ with respect to θ. The asymptotic variance of an asymptotically linear estimator is $E\psi\psi'$, which is apparently larger than that of the maximum likelihood estimator $\beta\left(\hat{\theta}\right)$ of any parametric sub-model, which is given through information matrix and the delta method as

$$\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)\left(E\left(S_\theta S_\theta'\right)\right)^{-1}\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)' = E\left[\psi S_\theta'\right]\left(ES_\theta S_\theta'\right)^{-1}E[S_\theta \psi'].$$

A starting point for calculating the semiparametric efficiency bound is to restrict attention to differentiable parameters $\beta(\theta)$ which satisfies $\frac{\partial \beta(\theta)}{\partial \theta} = E\left(dS_\theta'\right)$ for some d and all parametric sub-models. Such d are not unique. Adding a random vector that is orthogonal to S_θ preserves the validity of d. In fact, any linear influence function ψ can serve as a d. For differentiable parameters, if we use the invariance principle and the delta method, the Cramer–Rao lower bound for estimating $\beta(\theta)$ is

$$\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)\left(E\left(S_\theta S_\theta'\right)\right)^{-1}\left(\frac{\partial}{\partial \theta}\beta(\theta)\right)' = E\left[dS_\theta'\right]\left(ES_\theta S_\theta'\right)^{-1}E[S_\theta d']$$

Obviously, this is the variance of $d_\theta = E[dS_\theta](E[S_\theta S_\theta])^{-1}S_\theta$, which is the projection of d onto the linear space spanned by the score functions S_θ.

As the class of parametric sub-models expands, the linear space it spans also increases and the variance of d_θ also increases. The semiparametric efficiency bound should be the limit of this progress of increments. Formally, the tangent space is defined to be the mean square closure of all linear combinations of scores S_θ for smooth parametric sub-models, and the efficiency bound is given by the variance of the projection of d onto the tangent space T. In other words, the efficiency bound is given by $V = E[\delta\delta']$ where $\delta \in T$ and $E[d - \delta)'\iota] = 0$ for all $\iota \in T$.

Application

In this section we illustrate the computation of semiparametric efficiency bound using a model of non-classical measurement errors, studied in Chen et al. (2005) and Chen et al. (2004), where information from a primary data-set and from an auxiliary data-set need to be efficiently combined. Their models extend the results in the treatment effect literature on the mean parameter (see Hahn 1998, Hirano et al. 2003 and Imbens et al. 2005), to measurement error models where parameters

are generically defined through nonlinear moment conditions.

Consider the following model. The researcher is interested in a parameter β defined by the moment condition $Em(Y; \beta) = 0$ if and only if $\beta = \beta_0$. The researcher has access to a primary data-set which is a random sample from the population of interest. However, the true variable Y is not always observed in the primary data-set. Instead, a proxy variable X is observed throughout the primary data. For a subset of the primary data-set, which we will call the auxiliary data-set, X is validated so that both Y and X are observed. We will use the random variable $D = 0$ to denote observations in the auxiliary data-set where both X and Y are observed, and will use $D = 1$ to denote the rest of the primary data-set where only X is observed. Chen et al. (2004) call this the 'verify-in-sample' case. They make the following conditional independence assumption:

Assumption 4.1 $Y \perp D|X$.

Under this assumption, we follow the framework of Newey (1990) to show that the efficiency bound for estimating β is given by $\left(J_\beta \Omega_\beta^{-1} J_\beta\right)^{-1}$, where for

$$\mathcal{J}_\beta = \frac{\partial}{\partial \beta} E[m(Y;\beta)] \text{ and } \Omega_\beta = E\left[\frac{1}{1-p(X)} V[m(Y;\beta)|X] + \mathrm{E}(X;\beta)\mathrm{E}(X;\beta)'\right].$$

To demonstrate this result, we follow the steps in the efficiency framework of Newey (1990). First we characterize the properties of the tangent space under Assumption 4.1. Next we write the parameter of interest in its differential form and therefore find a linear influence function d. Finally, we conjecture and verify the projection of d onto the tangent space and the variance of this projection gives rise to the efficiency bound. We first go through these three steps under the assumption that the moment conditions exactly identify β. Finally, the results are extended to over-identified moment conditions by considering their optimal linear combinations.

First we assume that the moment conditions exactly identify β.

Step 1 Consider a parametric path θ of the joint distribution of Y, X and D. Define $p_\theta(x) = P_\theta(D = 1|\, x)$. Under assumption 1, the joint density function for Y, D and X can be factorized into

$$f_\theta(y,x,d) = f_\theta(x) p_\theta(x)^d [1 - p_\theta(x)]^{1-d} f_\theta(y|x)^{1-d}. \tag{1}$$

The resulting score function is then given by

$$S_\theta(d,y,x) = (1-d) s_\theta(y|x) + \frac{d - p_\theta(x)}{p_\theta(x)(1-p_\theta(x))} \dot{p}_\theta(x) + t_\theta(x),$$

where

$$s_\theta(y|x) = \frac{\partial}{\partial \theta} \log f_\theta(y|x), \dot{p}_\theta(x) = \frac{\partial}{\partial \theta} p_\theta(x), t_\theta(x) = \frac{\partial}{\partial \theta} \log f_\theta(x).$$

The tangent space of this model is therefore given by:

$$\mathcal{T} = \{(1-d) s_\theta(y|x) + a(x)(d - p_\theta(x)) + t_\theta(x)\} \tag{2}$$

where $\int s_\theta(y|x) f_\theta(y|x) dy = 0, \int t_\theta(x) f_\theta(x) dx = 0$, and $a(x)$ is any square integrable function.

Step 2 As in the method of moment model in Newey (1990), the differential form of the parameter β can be written as

$$\begin{aligned} \frac{\partial \beta(\theta)}{\partial \theta} &= -\left(\mathcal{J}_\beta\right)^{-1} E\left[m(Y;\beta) \frac{\partial \log f_\theta(Y,X)}{\partial \theta'}\right] \\ &= -\left(\mathcal{J}_\beta\right)^{-1} \{E[m(Y;\beta)(s_\theta(Y|X)' + t_\theta(X)')]\} \\ &= -\left(\mathcal{J}_\beta\right)^{-1} \{E[m(Y;\beta) s_\theta(Y|X)'] + E[\mathcal{E}(X) t_\theta(X)']\} \end{aligned} \tag{3}$$

Therefore $d = \mathcal{J}_\beta^{-1} m(Y;\beta)$. Since $\mathcal{J}_\beta$ is only a constant matrix of nonsingular transformation. The projection of d onto the tangent space will be $\mathcal{J}_\beta$

multiplied by the projection of $m(Y;\beta)$ onto the tangent space. Therefore we only need to consider the projection of $m(Y;\beta)$ onto the tangent space.

Step 3 We conjecture that this projection takes the form of

$$\tau(Y,X,D) = \frac{1-D}{1-p(X)}[m(Y;\beta) - \mathcal{E}(X)] + \mathcal{E}(X)$$

To verify that this is the efficient influence function we need to check that $\tau(Y, X,D)$ lies in the tangent space and that

$$E[(m(Y;\beta) - \tau(Y,X,D))s_\theta(Y,X)] = 0.$$

or that

$$E[m(Y;\beta)s_\theta(Y,X)] = E[\tau(Y,X,D)s_\theta(Y,X)]. \quad (4)$$

To see that $\tau(Y, X, D)$ lies in the tangent space, note that the first term in $\tau(Y, X, D)$ has mean zero conditional on X, and corresponds to the first term of $(1 - d)s_\theta(y|x)$ in the tangent space. The second term in $\tau(Y, X, D)$,$\mathsf{E}(x)$, has unconditional mean zero and obviously corresponds to the $t_\theta(x)$ in the tangent space.

To verify (4), one can make use of the representation of $E[m(Y;\beta)s_\theta(Y, X)]$ in (3), by verifying the two terms in $\tau(Y, X, D)$ separately. The second term is obvious and tautological. The first part,

$$E\left[\frac{1-D}{1-p(X)}[m(Y;\beta) - E(X)]s_\theta(Y,X)\right] = E[m(Y;\beta)s_\theta(Y,X)],$$

follows from the conditional independence Assumption 4.1 and the score function property $E[s\theta(Y, X)|\, X] = 0$. Therefore we have verified that $\tau(Y, X, D)$ is the efficient projection and that the efficiency bound is given by

$$V = \left(\mathcal{J}_\beta\right)^{-1} E\left[\tau(Y,X,D)\tau\left(Y,X,D\right)'\right]\left(J_\beta\right)^{\prime-1} = \left(\mathcal{J}_\beta\right)^{-1} E\left[\frac{1}{1-p(X)} Var\left(m(Y;\beta)\middle|X\right) + \mathcal{E}(X)\mathcal{E}(X)'\right]\left(\mathcal{J}_\beta\right)^{\prime-1}$$

Finally, consider the extensions of these results to the over-identified case. When $dm > d_\beta$, the moment condition is equivalent to the requirement that for any matrix A of dimension $d_\beta \times d_m$ the following exactly identified system of moment conditions holds

$$\mathsf{A}E(m(Y;\beta)] = 0.$$

Differentiating under the integral again, we have

$$\frac{\partial\beta(\theta)}{\partial\theta} = -\left(\mathsf{A}E\left[\frac{\partial m(Y;\beta)}{\partial\beta}\right]\right)^{-1} \times E\left[\mathsf{A}m(Y;\beta)\frac{\partial \log f_\theta\left(Y,X\middle|D=1\right)}{\partial\theta'}\right].$$

Therefore, any regular estimator for β will be asymptotically linear with influence function of the form

$$-\left(\mathsf{A}E\left[\frac{\partial m(Y;\beta)}{\partial\beta}\right]\right)^{-1}\mathsf{A}m(Y;\beta).$$

For a given matrix A, the projection of the above influence function onto the tangent set follows from the previous calculations, and is given by

$$-\left[\mathsf{A}J_\beta\right]^{-1}\mathsf{A}\tau(y,x,d).$$

The asymptotic variance corresponding to this efficient influence function for fixed A is therefore

$$\left[\mathsf{A}J_\beta\right]^{-1}\mathsf{A}\Omega\mathsf{A}'\left[J_\beta\mathsf{A}'\right]^{-1} \quad (5)$$

where

$$\Omega = E[\tau(Y,X,D)\tau(Y,X,D)']$$

as calculated above. Therefore, the efficient influence function is obtained when A is chosen to minimize this efficient variance. It is easy to show that the optimal choice of A is equal to $J'_\beta\Omega^{-1}$, so that the asymptotic variance becomes

$$V = \left(\mathcal{J}'_\beta\Omega^{-1}\mathcal{J}_\beta\right)^{-1}.$$

Different estimation methods can be used to achieve this semiparametric efficiency bound. In particular, Chen et al. (2004) showed that both a semiparametric conditional expectation projection estimator and a semiparametric propensity score estimator based on a sieve nonparametric first-stage regression achieve this efficiency bound.

Conclusion

As discussed in Newey (1990), while the calculation of the tangent space and the efficient projection is easy in several important examples, including the one above, it can be difficult in general. A variety of techniques are available to characterize the tangent space and the efficient projection. Some of these are discussed in details in Newey (1990) and Bickel et al. (1993).

Even in parametric models, the notion of asymptotic efficiency is more complex when one compares estimators that do not converge $\sqrt{n}$ rate or are not asymptotically distributed. Comparing these estimators requires the choice of a loss function, and different loss functions can lead to different efficiency rankings (see Ibragimov and Has'minskii 1981). In econometrics, these estimators sometimes arise in structural models in labour economics and in industrial organization. The efficiency properties of these estimators are analysed in Hirano and Porter (2003) and Chernozhukov and Hong (2004).

See Also

- ▶ Generalized Method of Moments Estimation
- ▶ Maximum Likelihood
- ▶ Measurement Error Models
- ▶ Non-parametric Structural Models
- ▶ Semiparametric Estimation
- ▶ Stratification

Bibliography

Amemiya, T. 1985. *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Amemiya, T. 1994. *Introduction to statistics and econometrics*. Cambridge, MA: Harvard University Press.

Bickel, P., C.A. Klaassen, Y. Ritov, and J. Wellner. 1993. *Efficient and adaptive estimation for semiparametric models*. New York: Springer-Verlag.

Chen, X., H. Hong, and A. Tarozzi. 2004. *Semiparametric efficiency in GMM models of nonclassical measurement errors*. Working paper: Duke University and New York University.

Chen, X., H. Hong, and E. Tamer. 2005. Measurement error models with auxiliary data. *Review of Economic Studies* 72: 343–366.

Chernozhukov, V., and H. Hong. 2004. Likelihood inference for a class of nonregular econometric models. *Econometrica* 72: 1445–1480.

Hahn, J. 1998. On the Role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66: 315–332.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054.

Hirano, K., and J. Porter. 2003. Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* 71: 1307–1338.

Hirano, K., G. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71: 1161–1189.

Ibragimov, I., and R. Has'minskii. 1981. *Statistical estimation: Asymptotic theory*. New York: Springer.

Imbens, G., W. Newey, and G. Ridder. 2005. Mean-squared-error calculations for average treatment effects. Working paper.

Newey, W. 1990. Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5(2): 99–135.

Newey, W., and D. McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of econometrics*, ed. R. Engle and D. McFadden, Vol. 4. Amsterdam: North-Holland.

Newey, W., and J. Powell. 1990. Efficient estimation of linear and type in censored regression models under conditional quantile restrictions. *Econometric Theory* 6: 295–317.

Powell, J. 1994. Estimation of semiparametric models. In *Handbook of econometrics*, ed. R. Engle and D. McFadden, Vol. 4. Amsterdam: North-Holland.

van der Vaart, A. 1999. *Asymptotic statistics*. Cambridge: Cambridge University Press.

Efficiency Wages

Andrew Weiss

Abstract

Efficiency wages capture the effect of compensation on the behaviour of workers, as well as on the quality of workers attracted and retained

by the firm. This effect has greater significance in some areas than others, and can be used to explain wage differentials among firms and industries, as well as to explain why firms respond to demand shocks by reducing their labour force rather than cutting wages, and may ration jobs even in normal times. At the macroeconomic level efficiency wages can explain persistent long-term unemployment as an equilibrium outcome in a competitive labour market.

Keywords

Capital cost; Capital market imperfections; Efficiency wages; Firm size; Firm-specific human capital; Gift exchange; Involuntary unemployment; Labour supply; Layoffs; Low-wage probation period; Monitoring; Nutrition models; Productivity; Retirement; Sorting effect of wages; Wage differentials

JEL Classifications

J310

'Efficiency wages' is a term used to express the idea that labour costs can be described in terms of efficiency units of labour rather than in terms of hours worked, and that wages affect the performance of workers. In this respect, labour differs from most other inputs (with the notable exception of credit), in which inputs are well defined independently of prices. Models of efficiency wages explore the implications of the interconnections between compensation and productivity. On the macroeconomic level, efficiency wages can explain persistent unemployment without relying on either structural imperfections such as search costs or fixed-length contracts or irrational behaviour such as money illusion, which would cause real wages to fail to adjust to market conditions. (For some of the earliest such models, see Futia 1977; Salop 1979; Solow 1979; Shapiro and Stiglitz 1984; Weiss 1981.) At the level of the firm, efficiency wages can result in job queues (excess supply of labour) and can explain why seemingly identical workers may receive different wages at different firms, and why these observed wage differentials are positively correlated with firm characteristics such as profitability, high capital–labour ratios, and establishment size (Brown and Medoff 1989). These market imperfections arise because employers cannot costlessly observe the ability and productivity of workers or because of capital market imperfections that prevent workers from 'buying' the high-wage jobs.

Efficiency wage models have one or more of the following characteristics:

1. Compensation levels and rules affect the types of workers who are attracted to, and retained by, the firm – this is normally referred to as the sorting effect of wages.
2. Compensation rules create incentives for workers to behave in ways that increase firm profits.
3. Wages affect the nutrition and health of workers and thus higher wages directly increase productivity (these 'nutrition' models are most applicable in poor countries).

Consequences of the use of efficiency wages are:

1. Compensation levels within a firm may not be proportionate to relative productivity.
2. Compensation could be a function of characteristics of the establishment employing the worker.
3. Wages could rise more steeply with tenure than does productivity.
4. Some firms could have an excess supply of workers.
5. A frictionless economy could be in a long-run equilibrium with unemployment.

The sorting effects of wages enable a firm to benefit from private information that the employee knows about himself and that is either not available to the firm or would be costly for the firm to acquire. High compensation enables the firm to draw from a larger and better pool of workers. Firms that test job applicants will also find that, by offering a higher wage, the expected

quality of the worker hired, conditional on the applicants test score, will also be higher.

The test could be in the form of a low-wage probation period for new hires. Using a low-wage probation period, followed by a significant wage increase, followed by high wages for workers who perform well during the probation period, the firm can attract job applicants with positive private information about their ability. If the test is imperfect, the use of a low-wage probation period will also discourage applications from risk-averse applicants as well as applicants with a higher cost of capital. Wages that increase steeply with tenure will attract workers who have low quit propensities (aside from their incentive effect of deterring quits). Groshen and Loh (1993) have found that much of the return to tenure takes place at the end of low-wage probationary periods.

Sorting effects of efficiency wages may also explain why firms do not cut wages in response to a fall in demand. If a firm were to cut the wages, it may find that its better workers are most likely to quit. Thus, a profit maximizing firm could find that its best response to a fall in demand for its product would be to fire workers rather than to cut wages.

Most of the efficiency wage models have focused on the ways in which compensation affects the behaviour of workers.

The incentive effects of wages stem from the effect of the level of compensation on the cost to the worker of being fired. Thus, wages above the market clearing level will increase effort, decrease employee theft, decrease absenteeism, and decrease quits. See, for example, Salop and Salop (1976), Klein et al. (1991), and Weiss (1984) on quits; Shapiro and Stiglitz (1984) on effort; Lazear and Rosen (1981), Weiss (1985) on absenteeism.

Levels of compensation also affect the attitude of the employee towards the firm. Thus, paying wages above the market clearing level may have multiple beneficial effects for the firm including: reducing employee theft, increasing unobserved effort, and inducing higher levels of care, which will decrease costs incurred from damage to the firm's property. Greater loyalty to the firm will also encourage workers to acquire firm-specific human capital, to report theft of firm property, and to allocate the worker's effort in ways that benefit the firm. See Akerlof (1984) on gift exchange.

Higher levels of compensation will also reduce the time needed to fill vacancies (Lang 1991). In this case the behaviour being affected is the application process.

Wages directly affect the productivity of workers through their effect on the nutrition of workers as well as their access to clean water and medical care and other goods and services that directly improve their productivity. These 'nutrition' effects are strongest in poor countries and could also possibly explain poverty traps for particularly poor workers who do not have access to firms that are offering efficiency wages.

The importance of these effects will vary across firms. For instance, we would expect that capital-intensive firms will derive the greatest benefit from reductions in absenteeism and quits, and from increased productivity of their employees. Capital-intensive firms will also tend to be most vulnerable to careless behaviour by workers that would damage the valuable property. Larger firms have more difficulty monitoring individual effort and directing the effort in ways that fit the needs of the firm. Consequently, the efficiency wage models would predict that compensation would be correlated with firm size. The direct effects of wages through better nutrition and health take some time to affect productivity, so we would expect that firms with lower costs of capital will offer higher wages – in poor countries these tend to be foreign firms. (In poor countries, in which the nutrition effects are strongest, we might see that wages would be correlated with a firm's cost of capital as well as with the ability of the firm to retain workers after their productivity has been enhanced by the higher wages. The nutrition effects of wages may take some time to affect productivity.) Finally, if high wages are used to attract better workers, then we would expect that when workers are laid off from firms in high-wage industries they will tend to get jobs in other high-wage industries (see Gibbons and Katz 1992).

All of these implications of the efficiency wage model have been confirmed by empirical studies

of the relationship between firm characteristics and wages. (In cases in which wages directly affects productivity we would expect that firms that are likely to be able to retain their workers will also pay higher wages. However, since wages directly affects turnover, and prices vary according to the presence of competitive firms, this implication of the nutrition version of the efficiency wage model is more difficult to verify.) Of course, many if not all of these empirical findings can be explained by other models. For example, the relationship between prior and posterior industry wages for laid-off workers can be explained by competitive models in which workers are being selected based on attributes, such as pulchritude, that are directly observed by the firm but not by the researchers.

Thus, efficiency wages can explain why empirical studies of the relationship between wage and characteristics of establishments find that large, capital-intensive establishments are most likely to pay wages that are above market clearing levels – and in the case of poor countries why foreign firms tend to pay higher wages. The efficiency wage models also can explain why firms fire workers rather than cutting wages, offer wages that attract an excess supply of workers, and pay some of their workers to take early retirement or seek to impose mandatory retirement. See, for instance, Brown and Medoff (1989). Finally, efficiency wage theory can explain the persistence of involuntary unemployment in a free market economy.

Bibliography

Akerlof, G.A. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97: 543–569.

Akerlof, G.A. 1984. Gift exchange and efficiency-wage theory: Four views. *American Economic Review* 74(2): 79–83.

Brown, C., and J. Medoff. 1989. The employer size-wage effect. *Journal of Political Economy* 97: 1027–1059.

Futia, C. 1977. Excess supply equilibria. *Journal of Economic Theory* 14: 200–220.

Gibbons, R., and L. Katz. 1992. Does unmeasured ability explain inter-industry wage differentials? *Review of Economic Studies* 59: 515–535.

Groshen, E., and E.S. Loh. 1993. What do we know about probationary periods? In *Proceedings of the 45th Annual Meeting of the Industrial Relations Research Association*, Madison.

Klein, R., R. Spady, and A. Weiss. 1991. Factors affecting the output and quit propensities of production workers. *Review of Economic Studies* 58: 929–954.

Landau, H., and A. Weiss. 1984. Wages, hiring standards and firm size. *Journal of Labor Economics* 2: 477–499.

Lang, K. 1991. Persistent wage dispersion and involuntary unemployment. *Quarterly Journal of Economics* 106: 181–202.

Lazear, E., and S. Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.

Salop, J., and S. Salop. 1976. Self-selection and turnover in the labor market. *Quarterly Journal of Economics* 90: 619–627.

Salop, S. 1979. A model of the natural rate of unemployment. *American Economic Review* 69: 117–125.

Shapiro, C., and J. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74: 433–444.

Solow, R. 1979. Another possible source of wage stickiness. *Journal of Macroeconomics* 1: 79–82.

Weiss, A. 1981. Job queues and layoffs in labor markets with flexible wages. *Journal of Political Economy* 88: 526–538.

Weiss, A. 1984. Determinants of quit behavior. *Journal of Labor Economics* 2: 371–387.

Weiss, A. 1985. Absenteeism and wages. *Economics Letters* 19: 277–279.

Weiss, A., and R. Wang. 1998. Probation, layoffs, and wage-tenure profiles: A sorting explanation. *Labour Economics* 5: 359–383.

Efficient Allocation

Stanley Reiter

JEL Classifications
D5

Analysis of efficiency in the context of resource allocation has been a central concern of economic theory from ancient times, and is an essential element of modern microeconomic theory. The ends of economic action are seen to be the satisfaction of human wants through the provision of goods and services. These are supplied by

production and exchange and limited by scarcity of resources and technology. In this context efficiency means going as far as possible in the satisfaction of wants within resource and technological constraints. This is expressed by the concept of Pareto optimality, which can be stated informally as follows: a state of affairs is Pareto optimal if it is within the given constraints and it is not the case that everyone can be made better off in his own view by changing to another state of affairs that satisfies the applicable constraints.

Because knowledge about wants, resources and technology is dispersed, efficient outcomes can be achieved only by coordination of economic activity. Hayek (1945) pointed out the role of knowledge or information, particularly in the context of prices and markets, in coordinating economic activity. Acquiring, processing and transmitting information are costly activities themselves subject to constraints imposed by technological and resource limitations. Hayek pointed out that the institutions of markets and prices function to communicate information dispersed among economic agents so as to bring about coordinated economic action. He also drew attention to motivational properties of those institutions, or incentives. In this context, the concept of efficiency takes account of the organizational constraints on information processing and transmission in addition to those on production of ordinary goods and services. The magnitude of resources devoted to business or governmental bureaucracies, and to some of the functions performed by industrial salesmen, attests to the importance of these constraints. Economic analysis of efficient allocation has formally imposed only the constraints on production and exchange, and until recently recognized organizational constraints only in an informal way. But it is these constraints that motivate the pervasive and enduring interest in decentralized modes of economic organization, particularly the competitive mechanism.

It is necessary to limit the scope of this essay so that it is not coextensive with microeconomic theory. The main limitation imposed here is to confine attention to models in which either the role of information is ignored, or in which agents do not behave strategically on the basis of private information. In so doing, a large and important class of models involving problems of efficient allocation in the presence of incentive constraints is excluded.

The main ideas of efficient resource allocation are present in their simplest form in the linear activity analysis model of production. We begin with that model.

E

Efficiency of Production: Linear Activity Analysis

The analysis of production can to some extent be separated from that of other economic activity. The concept of efficiency appropriate to this analysis descends from that of Pareto optimality, which refers to both productive and allocative efficiency in the full economy in which production is embedded. It is useful to begin with a model in which technological possibilities afford constant returns to scale, that is, with the (linear) activity analysis model of production pioneered by Koopmans (1951a, b, 1957), and closely related to the development of linear programming associated with Dantzig (1951a, b) and independently with the Russian mathematician Kantorovitch (1939, 1942) and Kantorovitch and Gavurin (1949).

The two primitive concepts of the model are *commodity* and *activity*. A list of n commodities is postulated; a commodity *bundle* is given by specifying a sequence of n numbers $a_1, a_2, \ldots, a_n$. Technological possibilities are thought of as knowledge of how to transform commodities. Such knowledge may be described in terms of collections of activities called *processes*, much as knowledge of how to prepare food is described by recipes. A recipe commonly has two parts, a list of ingredients or inputs and of the output(s) of the recipe, and a description of how the ingredients are to be combined to produce the output(s). In the activity analysis model the description of productive activity is suppressed. Only the specification of inputs and outputs is retained; this defines the production process.

Commodities are classified into 'desired', 'primary' and 'intermediate' commodities. Desired commodities are those whose consumption or availability is the recognized goal of production; they satisfy wants. Primary commodities are those available from nature. (A primary commodity that is also desired is listed separately among the desired commodities and must be transformed by an act of production into its desired form). Intermediate commodities are those that merely pass from one stage of production to another. Each commodity can exist in any non-negative amount (*divisibility*). Addition and subtraction of the numbers measuring the amount of a commodity represent joining and separating corresponding amounts of the commodity.

An activity is characterized by a *net output number* for each commodity, which is positive if the commodity is a net output, negative if it is a net input and zero if it is neither. The term *input-output vector* is also used for this ordered array of numbers. Activity analysis postulates a finite number of basic activities from which all technologically possible activities can be generated by suitable combination. Allowable combinations are as follows. If two activities are known to be possible, then the activity given by their algebraic sum is also possible, i.e. if $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$, then $a + b = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)$ is also possible. Thus, additivity embodies an assumption of non-interaction between productive activities, at least at the level of knowledge. Furthermore, if an activity is possible, then so is every non-negative multiple of it (*proportionality*), i.e. if $a = (a_1, a_2, \ldots, a_n)$ is possible, then so is $\mu a = (\mu a_1, \mu a_2, \ldots, \mu a_n)$ for any non-negative real number μ. This expresses the assumption of constant returns to scale. The family of activities consisting of all non-negative multiples of a given one forms a process. Since there is a finite number of basic activities, there is also a finite number of basic processes, each intended to describe a basic method of production capable of being carried out at different levels, or intensities.

The assumptions of additivity and proportionality determine a linear model of technology that can be given the following form. Let A be an n by k matrix whose jth column is the input-output vector representing the basic activity that defines the jth basic process, and let $x = (x_1, x_2, \ldots, x_n)$ be the vector whose jth component x_j is the scale (level or intensity) of the jth basic process. Let $y = (y_1, y_2, \ldots, y_n)$ be the vector of commodities. Technology is represented by a linear transformation mapping the space of activity levels into the commodity space, i.e.

$$y = Ax \quad x \geq 0.$$

With the properties assumed, a process can be represented geometrically in the commodity space by a halfline from the origin including all non-negative multiples of some activity in that process. The finite number of halflines representing basic processes generate a convex polyhedral cone consisting of all activities that can be expressed as sums of activities in the basic processes, or equivalently, as non-negative linear combinations of the basic activities, sometimes called a *bundle of basic activities*. This cone is called the *production set*, or set of *possible productions*.

Two other assumptions are made about the production set itself, rather than just the individual activities. First, there is no activity, whether basic or derived, in the production set with a positive net output of some commodity and non-negative net outputs of all commodities. This excludes the possibility of producing something from nothing, whether directly or indirectly. Second, it is assumed that the production set contains at least one activity with a positive net output of some commodity.

If the availability of primary commodities is subject to a bound, the technologically possible productions described by the production set are subject to another restriction; only those possible productions that do not require primary inputs in amounts exceeding the given bounds can be produced. Furthermore, because intermediate commodities are not desired in themselves, their net output is required to be zero. (Strictly speaking, the technological constraint on intermediate commodities is that their net output be non-negative. The requirement that they be zero can be viewed as one of elementary efficiency, excluding

accumulation or necessity to dispose of unwanted goods.) With these restrictions the model can be written

$$y = Ax, \quad x \geq 0, \quad y_i = 0$$

if i is an intermediate commodity, and

$$y_i \geq r_i$$

if i is a primary commodity, where r_i is the (non-positive) limit on the availability of primary commodity i. This leads to the concept of an *attainable* activity.

A bundle of basic activities is *attainable* if the resulting net outputs are non-negative for all desired commodities, zero for intermediate commodities and non-positive for primary commodities, and if the total inputs of primary commodities do not exceed (in absolute amount) the prescribed bounds of availability of those commodities. The set of activities satisfying these conditions is a truncated convex polyhedral cone in the commodity space called the *set of attainable productions*.

The concept of productive efficiency in this model is as follows. An activity (a bundle of basic activities) is *efficient* if it is attainable and if every activity that provides more of some desired commodity and no less of any other is not attainable.

This concept can be seen to be a specialization of Pareto optimality. If for each desired commodity there is at least one consumer who is not satiated in that commodity, at least in the range of production attainable within the given resource limitations, then increasing the amount of any desired commodity without decreasing any other can improve the state of some non-satiated consumer without worsening that of any other.

Characterizing Efficient Production in Terms of Prices

Efficient production can be characterized in terms of *implicit prices*, also called *shadow prices*, or in the context of linear programming, *dual variables*. Efficient activities are precisely those that maximize profit for suitably chosen prices. The profit returned by a process carried out at the level x is

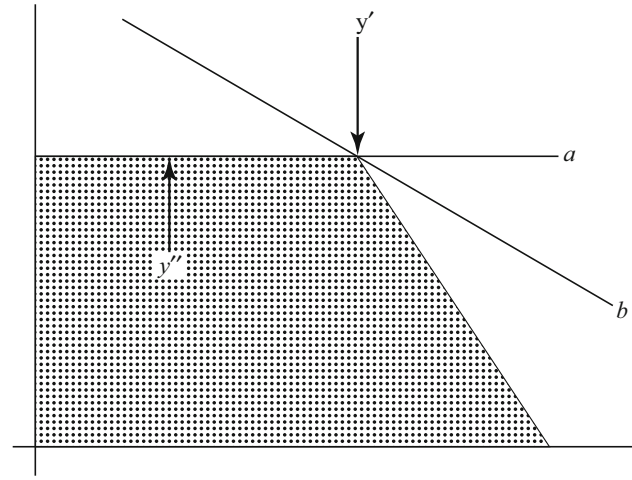
$$x \sum_i p_i a_i,$$

where the prices are $p = (p_1, \ldots, p_n)$, and $a = (a_1, \ldots, a_n)$ is the basic activity defining theprocess; the profit on the bundle of activities Ax at prices p is given by the inner product $py = pAx$.

This characterization is the economic expression of an important mathematical fact about convex sets in $n - 1$ dimensional Euclidean space, namely that through every point of the space not interior to the convex set in question there passes a hyperplane that contains the set in one of its two halfspaces (Fenchel 1950; Nikaido 1969, 1970). (A hyperplane in n dimensional space is a level set of a linear function of n variables, and thus is a translate of an $n - 1$ dimensional linear subspace. A hyperplane is given by an equation of the form $c_1x_1 + c_2x_2 + \cdots + c_nx_n = k$, where the x's are variables, the c's are coefficients defining the linear function and k is a constant identifying the level set. A hyperplane divides the space into two halfspaces corresponding to the two inequalities $c_1x_1 + c_2x_2 + \cdots + c_nx_n \gtreqless k$ respectively.) It can also be seen that a point of a convex set is a boundary point if and only if it maximizes a linear function on the (closure of the) set. These facts can be used to characterize efficient production because the attainable production set is convex and efficient activities are boundary points of it. Because the efficient points are those, roughly speaking, on the 'north-east' frontier of the set, the linear functions associated with them have non-negative coefficients, interpreted as prices. On the other hand, if a point of the attainable set maximizes a linear function with strictly positive coefficients (prices), then it is on the 'north-east' frontier of the set.

In Fig. 1 the set enclosed by the broken line and the axes is the projection of the attainable set on the output coordinates; inputs are not shown. The point y' in the figure is efficient; the point y' is not; both y' and y'' maximize a linear function with

Efficient Allocation, Fig. 1



non-negative coefficients (the level set containing y' is labelled a and also contains y''). However, y' maximizes a linear function with positive coefficients (one such, whose level set through y' is labelled b, is shown), while y'' does not.

These implicit, or efficiency prices arise from the logic of efficiency or maximization when the relevant sets are convex, not from any institutions such as markets or exchange. An important reason for interest in them is the possibility of achieving efficient performance by decentralized methods. As described above, under the assumptions of additivity and constant returns to scale the production set can be seen to be generated by a finite number of basic processes, each of which consists of the activities that are non-negative multiples of a basic activity, the multiple being the scale (level, or intensity) at which the process is operated. Following the presentation of Koopmans (1957), each basic process is controlled by a manager, who decides on its level. The manager of a process is assumed to know only the input-output coefficients of his process. Each primary resource is in the charge of a resource holder, who knows the limit of its availability. Efficiency prices are used to guide the choices of managers and resource holders. (Under constant returns to scale, if an activity yields positive profit at a given system of prices, then increasing the scale of the process containing that activity increases the profit. Since the scale can be increased without bound, if the profitability of a process is not zero or negative, then, in the eyes of its manager, who does not know the aggregate resource constraints, it can be made infinite. Therefore, the systems of prices that can be considered for the role of efficiency prices must be restricted to those *compatible with the given technology*, namely prices such that no process is profitable and at least one process breaks even). Two propositions characterize efficient production by prices and provide the basis for an interpretation in terms of decentralized control of production.

> In a given linear activity analysis model, if there is a given system of prices compatible with the technology, in which the prices of all desired commodities are positive, then any attainable bundle of basic activities selected only from processes that break even and which utilizes all positively priced primary commodities to the limit of their availability and does not use negatively priced primary commodities at all, is an efficient bundle of activities.
>
> In a given linear activity analysis model, each efficient bundle of activities has associated with it at least one system of prices compatible with the technology such that every activity in that bundle breaks even and such that prices of desired commodities are positive, and the price of a primary commodity is non-negative, zero or non-positive, according as its available supply is full, partly, or not used at all (Koopmans 1957).

These propositions are stated in a static form. There is no reference to managers raising or lowering the levels of the processes they control, or to

resource holders adjusting prices. A dynamic counterpart of these propositions would be of interest, but because of the linearity of the model such dynamic adjustments are unstable (Samuelson 1949).

It should also be noted that the concept of decentralization is not explicitly defined in this literature; the interpretation is by analogy with the competitive mechanism. Nevertheless, the interest in characterizing efficiency by prices and their interpretation in terms of decentralization is an important theme in the study of efficient resource allocation.

The linear activity analysis model has been generalized in several directions. These include dropping the assumption of proportionality, dropping the restriction to a finite number of basic activities, dropping the restriction to a finite number of commodities and dropping the restriction to a finite number of agents. Perhaps the most directly related generalization is to the nonlinear activity analysis, or nonlinear programming, model.

Efficiency of Production: Nonlinear Programming

In the nonlinear programming model there is, as in the linear model, a finite number of basic processes. Their levels are represented by a vector $x = (x_1, x_2, \ldots, x_k)$, where k is the number of basic processes. Technology is represented by a nonlinear transformation from the space of process levels to the commodity space (still assumed to be finite dimensional), written

$$y = F(x), \quad x \geq 0.$$

The production set in this model is the image in the commodity space of the non-negative orthant of the space of process levels. Under the assumptions usually made about F, the production set is convex, though, of course, not a polyhedral cone.

In this model as in the linear activity analysis model a central result is the characterization of efficient production in terms of prices. The simplest case to begin with is that of one desired commodity, say, one output, with perhaps several inputs. In this case the (vector-valued) function F can be written

$$F(x) = [f(x), g_1(x), g_2(x), \ldots, g_m(x)],$$

where the value of f is the output, and $g_1, \ldots, g_m$ correspond to the various inputs. Resource constraints are expressed by the conditions

$$g_j(x) \geq 0, \quad for\ j = 1, 2, \ldots, m,$$

and non-negativity of process levels by the condition, $x \geq 0$. (Here the resource constraints $r_j \leq h_j(x) \leq 0$ are written more compactly as $h_j(x) - r_j = g_j(x) \geq 0$.)

In this model the definition of efficient production given in the linear model amounts to maximizing the value of f subject to the resource and non-negativity constraints just mentioned.

Problems of constrained maximization are intimately related to saddle-point problems. Let L be a real valued function defined on the set $X \times Y$ in R^n. A point (x^*, y^*) in $X \times Y$ is a *saddle point* of L if

$$L(x, y^*) \leq L(x^*, y^*) \leq L(x^*, y),$$

for all x in X and all y in Y. The concept of a concave function is also needed. A real valued function f defined on a convex set X in R^n is a *concave function* if for all x and y in X and all real numbers $0 \leq a \leq 1$

$$f(ax + (1 - a)y) \geq af(x) + (1 - a)f(y).$$

The following mathematical theorem is fundamental.

Theorem (Kuhn and Tucker 1951; Uzawa 1958): Let f and $g_1, g_2, \ldots, g_m$ be real valued concave functions defined on a convex set X in R^n. If f achieves a maximum on X subject to $g_j(x) \geq 0$, $j = 1, 2, \ldots, m$ at the point x^* in X, then there exist non-negative numbers $p_0^*, p_0^*, \ldots, p_m^*$, not all zero, such that $p_0^* f(x) + p^* g(x) \leq p_0^* f(x^*)$ for all x in X, and furthermore, $p^* g(x)^* = 0$. (Here the vectors $p^* = (p_1^*, p_2^*, \ldots, p_m^*)$, and $g(x) = [g_1(x), g_2(x), \ldots, g_m(x)]$). The vector p^* may be chosen so that

$$\sum_{0}^{m} p_j^* = 1.$$

An additional condition (Slater 1950) is important. (It ensures that the coefficient p_0 of f is not zero.)

Slater's Condition There is a point x' in X at which $g_j(x') > 0$ for all $j = 1, 2, \ldots, m$.

If attention is restricted to concave functions, as in the Kuhn-Tucker-Uzawa Theorem, the relation between constrained maxima and saddle points can be summarized in the following theorem.

Theorem If f and $g_j, j = 1, 2, \ldots, m$ are concave functions defined on a convex subset X in R^n, and if Slater's Condition is satisfied, then x^* in X maximizes f subject to $g_j(x) \geq 0, j = 1, 2, \ldots, m$, if and only if there exists $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*)$, $\lambda_j^* \geq 0$ for $j = 1, 2, \ldots, m$, such that (x^*, λ^*) is a saddle point of $L(x, \lambda) = f(x) + \lambda g(x)$ on $X \times R_+^n$.

This theorem is easily seen to cover the case where some constraints are equalities, as in the case of intermediate commodities. The sufficiency half of this theorem holds for functions that are not concave.

The auxiliary variables $\lambda_1, \lambda_2, \ldots, \lambda_m$, called *Lagrange multipliers*, play the role of efficiency prices, or shadow prices; they evaluate the resources constrained by the condition $g(x) \geq 0$. The maximum characterized by the theorem is a global one, as in the case of linear activity analysis.

If the functions involved are differentiable, a saddle point of the Lagrangean can be studied in terms of first-order conditions. The first-order conditions are necessary conditions for a saddle point of L. If the functions f and the g's are concave on a convex set X, then the first-order conditions at a point (x^*, λ^*) are also sufficient; that is, they imply that (x^*, λ^*) is a saddle point of L. Thus,

Theorem If $f, g_1, g_2, \ldots, g_m$ are concave and differentiable on an open convex set X in R^n, and if Slater's Condition is satisfied, then x^* maximizes f subject to $g_j(x) \geq 0$ for $j = 1, 2, \ldots, m$ if and only if there exists numbers $\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*$ such that the first-order conditions for a saddle point of $L(x, \lambda) = f(x) + \lambda g(x)$ are satisfied at (x^*, λ^*).

If there are non-negativity conditions on the x's,

$$g_j(x) \geq 0, \quad x \geq 0, \quad x \text{ in } R^n$$

and the first-order conditions can be written

$$\begin{gathered} f_x^* + \lambda^* g_x^* \leq 0, \left(f_x^* + \lambda^* g_x^*\right) x^* = 0, \\ \lambda^* g(x^*) = 0, \quad g(x^*) \geq 0, g(x^*) \geq 0, \\ \lambda^* \geq 0 \quad \text{and } \lambda^* g(x^*) = 0, \end{gathered}$$

where f_x^* denotes the derivative of f evaluated at x^*. In more explicit notation, the conditions $f_x^* + \lambda^* g_x^* = 0$ can be written as

$$\partial f / \partial x_i + \sum_{j=1}^{m} \lambda_j^* \partial g_j / \partial x_i = 0, \quad i = 1, 2, \ldots, n$$

When the assumption of concavity is dropped, it is no longer possible to ensure that the local maximum is also a global one. However, it is still possible to analyse local constrained maxima in terms of local saddle-point conditions. In this case a condition is needed to ensure that the first-order conditions for a saddle point are indeed necessary conditions. The Kuhn-Tucker Constraint Qualification is such a condition. Arrow et al. (1961) have found a number of conditions, more useful in application to economic models, that imply the Constraint Qualification.

The case of more than one desired commodity leads to what is called the *vector maximum problem*, Kuhn and Tucker (1951). This may be defined as follows. Let $f_1, f_2, \ldots, f_k$ and $g_1, g_2, \ldots, g_m$ be real valued functions defined on a set X in R^n. We say x^* in X achieves a (global) *vector maximum* of $f = (f_1, f_2, \ldots, f_k)$ subject to $g_j(x) \geq 0, j = 1, 2, \ldots, m$ if,

(I) $g_j(x^*) \geq 0, j = 1, 2, \ldots, m$,

(II) there does not exist x' in X satisfying $f_i(x') \geq f_i(x^*)$ for $i = 1, 2, \ldots, k$ with $f_i(x') \geq f_i(x^*)$ for some value of i, and $g_j(x') \geq 0$ for $j = 1, 2, \ldots, m$.

This is just the concept of an efficient point expressed in the present notation.

A vector maximum has a saddle-point characterization similar to that for a scalar valued function.

Theorem Let $f_1, f_2, \ldots, f_k$ and $g_1, g_2, \ldots, g_m$ be real valued concave functions defined on a convex X set in R^n. Suppose there is x^0 in X such that $g_j(x^0) \geq 0, j = 1, 2, \ldots, m$, (Slater's Condition). If x^* achieves a vector maximum of f subject to $g(x) \geq 0$ then there exist $a = (a_1, a_2, \ldots, a_k)$ and $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*)$ with $a_j \geq 0$ for all j, $a \neq 0$ and $\lambda \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrangean $L(x, \lambda) = af(x) + \lambda g(x)$.

Several different 'converses', to this theorem are known. One states that if x^* maximizes $L(x, \lambda^*)$ for some strictly positive vector a and non-negative λ^*, and if $\lambda^* g(x^*) = 0$ and $g(x^*) \geq 0$, then x^* gives a vector maximum of f subject to $g(x) \geq 0$, and x in X. Another, parallel to the result for the case of one desired commodity, is the following.

Theorem Let f and g be functions as in the theorem above. If there are positive real numbers $a_1, a_2, \ldots, a_k$ and if (z^*, λ^*) is a saddle point of the Lagrangean L (defined as above) then (I) x^* achieves a maximum of f subject to $g(x) \geq 0$ on X, and (II) $\lambda^* g(x^*) = 0$.

The positive numbers $a_1, \ldots, a_k$ are interpreted as prices of desired commodities, and the non-negative numbers λ_j^* are prices of the remaining commodities. The condition $\lambda^* g(x^*) = 0$ which arises in these theorems states that the value of unused resources at the efficiency prices λ^* is zero; that is, resources not fully utilized at a vector maximum have a zero price.

The connection between vector maxima and Pareto optima is as follows. Because a vector maximum is an efficient point (for the vectorial ordering of the commodity space), it is a Pareto optimum for appropriately specified (non-satiated) utility functions, as was already pointed out in the case of the linear activity analysis model. Furthermore, if the functions $f_1, \ldots, f_k$ are themselves utility functions, and the variable x denotes allocations, with the constraints g defining feasibility, then a vector maximum of f subject to the constraints $g(x) \geq 0$ and x in X is a Pareto optimum, and vice versa. Hence the saddle-point theorems give a characterization of Pareto optima by prices. The interpretation of prices in terms of decentralized resource allocation described in the linear activity analysis model also applies in this nonlinear model. The proofs of these theorems reveal an important logical role played by the principle of marginal cost pricing.

The basic theorems of nonlinear programming, especially the Kuhn-Tucker-Uzawa Theorem in the setting of the vector maximum problem, have been extended to the case of infinitely many commodities. (Hurwicz 1958, first obtained the basic results in this field.) Technicalities aside, the theorems carry over to certain infinite dimensional spaces, namely linear topological spaces, or in the case of first-order conditions, Banach spaces.

Dropping the restriction to a finite number of basic processes leads to classical production or transformation function models of production, whose properties depend on the detailed specifications made.

Samuelson (1947) used Lagrangean methods to analyse interior maxima subject to equality constraints in the context of production function models, as well as that of optimization by consumers. He also gave the interpretation of Lagrange multipliers as shadow prices.

Efficient Allocation in an Economy with Consumers and Producers

In an economy with both consumption and production decisions, efficiency is concerned with distribution as well as production. Data about restrictions on consumption and the wants of consumers must be specified in addition to the data about production. The elements of the models are as follows.

The commodity space is denoted X; it might be l-dimensional Euclidean space, or a more abstract space such as an additive group in which, for example, some coordinates are restricted to have integer values. There is a (finite) list of consumers, 1, 2,. . ., n, and a similar list of producers, 1, 2,. . ., m. A *state* of the economy is an array consisting of a commodity bundle for

each agent in the economy, consumer or producer. This may be written $(\langle x^i\rangle, \langle y^j\rangle)$, where $\langle x^i\rangle = (x^1, x^2, \ldots, x^n)$ and $\langle y^j\rangle = (y^1, y^2, \ldots, y^m)$ and x^i and y^j are commodity bundles. Absolute constraints on consumption are expressed by requiring that the allocation $\langle x^i\rangle$ belong to a specified subset X of the space X^n of allocations.

Examples of such constraints are:

1. The requirement that the quantity of a certain commodity be non-negative.
2. The requirement that a consumer requires certain minimum quantities of commodities in order to survive.

Each consumer i has a preference relation, denoted $\succsim_i$, defined on X. This formulation admits externalities in consumption, including physical externalities and externalities in preferences; for example, preferences that depend on the consumption of other agents, termed non-selfish preferences. The consumption set of the ith consumer is the projection X^i of X onto the space of commodity bundles whose coordinates refer to the holdings of the ith consumer.

Technology is specified by a production set Y, a subset of X^m, consisting of those arrays $\langle y^j\rangle$ of input-output vectors that are jointly feasible for all producers. The production set of the jth producer, denoted Y^j, is the projection of Y onto the subspace of X^m whose coordinates refer to the jth producer.

The (aggregate) initial endowment of the economy is denoted by w, a commodity bundle in X.

These specifications define an *environment*, a term introduced by Hurwicz (1960) in this usage and according to him suggested by Jacob Marschak. This term refers to the primitive or given data from which analysis begins. Each environment determines a set of *feasible* states. These are the states $(\langle x^i\rangle, \langle y^j\rangle)$ such that $\langle x^i\rangle$ is in X, $\langle y^j\rangle$ is in Y and

$$\sum x^i - \sum y^j \leq w.$$

An environment determines the set of states that are Pareto optimal for that environment. Explicitly, they are the states $(\langle x^{*i}\rangle, \langle y^{*j}\rangle)$ that are feasible in the given environment, and such that if any other state $(\langle x^i\rangle, \langle y^j\rangle)$ has the property that $\langle x^i\rangle \succsim_i \langle x^{*i}\rangle$ for all i with $\langle x^i\rangle \succ_i \langle x^{*i}\rangle$ for some i', then $(\langle x^i\rangle, \langle y^j\rangle)$ is not feasible in the given environment.

It is important to note that the set of feasible states and the set of Pareto optimal states are completely determined by the environment; specification of economic organization is not involved.

At this level of generality, where externalities in consumption and production are admitted as possibilities, and where commodities may be indivisible, no general characterization of Pareto optima in terms of prices is possible. (Indeed, Pareto optima may not exist. Conditions that make the set of feasible allocations non-empty and compact and preferences continuous suffice to ensure the existence of Pareto optima.) In environments with externalities, or other non-neoclassical features, Pareto optima are generally not attainable by decentralized processes.

If the class of environments under consideration is restricted to the neoclassical environments, the fundamental theorems of welfare economics provide a characterization of Pareto optimal states via efficiency prices. That characterization has a natural interpretation in terms of a decentralized mechanism for allocation of resources.

The framework for these results is obtained by restricting the class of environments specified above as follows. The commodity space is to be Euclidean space of l dimensions, i.e. $X = R^l$. The consumption set for the economy is to be the product of its projections, i.e. $X = X^1 \times X^2 \times \cdots \times X^n$. This expresses the fact that if each agent's consumption is feasible for him, the total array is jointly feasible. Furthermore, each agent is restricted to having selfish preferences; that is, agent i's preference relation depends only on the coordinates of the allocation that refer to his holdings. In that case the preference relation $\succsim_i$ may be defined only on X^i, for each i. Similarly, externalities are ruled out in production, i.e. $Y = Y^1 \times Y^2 \times \cdots \times Y^m$.

The concept of an *equilibrium relative to a price system* (Debreu 1959) serves to characterize Pareto optima by prices. A price system, denoted p, is an

element of R^l; the environment $e = [(X^i), (\succsim_i), (Y^j), w]$ is of the restricted type specified above (free of externalities and indivisibilities).

A state $[(x^{*i}), (y^{*j})]$ of e is an *equilibrium relative to price system p* if:

1. For every consumer i, x^{*i} maximizes preference $\succsim_i$ on the set of consumption bundles whose value at the prices p does not exceed the value of x^{*i} at those prices, i.e., if x^i is in $\{x^i \text{ in } x^i : px^i \leq px^{*i}\}$ then $x^i \preccurlyeq ix^{*i}$.
2. For every producer j , y^{*j} maximizes profit py^j on Y^j.
3. Aggregate supply and demand balance, i.e.

$$\sum_i x^{*i} - \sum_i y^{*j} = w.$$

An equilibrium relative to a price system differs from a competitive equilibrium (see below) in that the former does not involve the budget constraints applying to consumers in the latter concept. In an equilibrium relative to a price system the distribution of initial endowment and of the profits of firms among consumers need not be specified.

The first theorem of neoclassical welfare economics states, subject only to the exclusion of externalities and a mild condition that excludes preferences with thick indifference sets, that a state of an environment e that is an equilibrium relative to a price system p is a Pareto optimum of e (Koopmans 1957).

The second welfare theorem is deeper and holds only on a smaller class of environments, sometimes referred to in the literature as the *classical environments* (called neoclassical above). One version of this theorem is as follows. Let $e = [(X^i), (\succsim_i), (Y^j), w]$ be an environment such that for each i

1. X^i is convex.
2. The preference relation $\succsim_i$ is continuous.
3. The preference relation $\succsim_i$ is convex.
4. The set $\Sigma_j Y_j$ is convex.

Let $[(x^{*i}), (y^{*j})]$ be a Pareto optimum of e such that there is at least one consumer who is not satiated at x^{*i}. Then there is a price system p, with not all components equal to 0, such that – except for Arrow's (1951) 'exceptional case', where p is such that for some i the expenditure px^{*i} is a minimum on the consumption set X^i – the state $[(x^{*i}), (y^{*j})]$ is an equilibrium relative to p.

(The condition that preferences are convex and not satiated is sufficient to exclude 'thick' indifference sets. A preference relation on X^i is convex if whenever x' and x'' are points of X^i with x' strictly preferred to x'' then the line segment connecting them (not including the point x'') is strictly preferred to x'. The consumption set X^i must be convex for this property to make sense. A preference relation is not satiated if there is no consumption preferred to all others.)

Hurwicz (1960) has given an alternative formalization of the competitive mechanism in which Arrow's exceptional case presents no difficulties.

If the exceptional case is not excluded, then it can still be said that:

1. x^{*i} minimizes expenditure at prices p on the upper contour set of x^{*i}, for every i, and
2. y^{*j} maximizes 'profit' py^j on the production set Y^j, for every j.

The state (x^*, y^*) together with the prices p, constitute a *valuation equilibrium* (Debreu 1954).

As in the case of efficiency prices in pure production models, these prices have in themselves no institutional significance. They are, however, in the same way as other efficiency prices, suggestive of an interpretation in terms of decentralization.

If, in addition to the restriction to classical environments, the economic organization is specified to be that of a system of markets in a private ownership economy, and if agents are assumed to take prices as given, then the welfare theorems can translate into the assertion that the set of Pareto optima of an environment e and the set of competitive equilibria for e (subject to the possible redistribution of initial endowment and ownership shares) are identical. More precisely, the specification of the environment given above is augmented by giving each consumer a bundle of commodities, his initial endowment, denoted w^i.

The total endowment is $w = \Sigma_i w^i$. Furthermore, each consumer has a claim to a share of the profits of each firm; the claims for the profit of each firm are assumed to add up to the entire profit. When prices and the production decisions of the firms are given, the profits of the firms are determined and so is the value of each consumer's initial endowment. Therefore, the income of each consumer is determined. Hence, the set of commodity bundles a consumer can afford to buy at the given prices, called his *budget set*, is determined; this consists of all bundles in his consumption set whose value at the given prices does not exceed his income at the given prices. Competitive behaviour of consumers means that each consumer treats the prices as given constants and chooses a bundle in his budget set that maximizes his preference: that is, a bundle x^i that is in X^i and such that if any other bundle x'^i is preferred to it, then x'^i is not in his budget set.

Competitive behaviour of firms is to maximize profits computed at the given prices p, regarded by the firms as constants; that is, a firm chooses a production vector y^j in its production set with the property that any other vector affording higher profits than py^j is not in the production set of firm j.

A *competitive equilibrium* is a specification of a commodity bundle for each consumer, a production vector for each firm, and a price system, together denoted $[(x^{*i}), (y^{*j}), p^*]$, where p^* has no negative components, satisfying the following conditions:

1. For each consumer i the bundle x^{*i} maximizes preference on the budget set of i.
2. For each firm j the production vector y^{*j} maximizes profit p^*y^j on the production set Y^j.
3. For each commodity, the total consumption does not exceed the net total output of all firms plus the total initial endowment, i.e. $\Sigma_i x^{*i} - \Sigma_j y^{*j} \leq w = \Sigma_i w^i$;
4. For those commodities k for which the inequality in 3 is strict; that is, the total consumption is less than initial endowment plus net output, the price p_k^* is zero.

The welfare theorems stated in terms of equilibrium relative to a price system translate directly into theorems stated in terms of competitive equilibrium. Briefly, every competitive equilibrium allocation in a given classical environment is Pareto optimal in that environment, and every Pareto optimal allocation in a given classical environment can be made a competitive equilibrium allocation of an environment that differs from the given one only in the distribution of the initial endowment. (Arrow (1951), Koopmans (1957), Debreu (1959) and Arrow and Hahn (1971) give modern and definitive treatment of the classical welfare theorems.)

It should be noted that the equilibria involved must exist for these theorems to have content. Sufficient conditions for existence of competitive equilibrium, which, since a competitive equilibrium is automatically an equilibrium relative to a price system, are also sufficient for existence of an equilibrium relative to a price system, include convexity and continuity of consumption sets and preferences and of production sets, as well as some assumptions which apply to the environment as a whole, restricting the ways in which individual agents may fit together to form an environment (Arrow and Debreu 1954; Debreu 1959; McKenzie 1959).

The second welfare theorem involves redistribution of initial endowment. This is essential because the set of competitive equilibria from a given initial endowment is small (essentially finite) (Debreu 1970), while the set of Pareto optima is generally a continuum. The set of Pareto optima cannot in general be generated as competitive allocations without varying the initial point. If redistribution is done by an economic mechanism, then it should be a decentralized one to support the interpretation given of the second welfare theorem. No such mechanism has been put forward as yet. Redistribution of initial endowment by lump-sum taxes and transfers has been discussed. A customary interpretation views these as brought about by a process outside economics, perhaps by a political process; no claim is made that such processes are decentralized. Some economists consider dependence on redistribution unsatisfactory because information about initial endowment is private; only the individual agent knows his own endowment. Consequently the

expression of that information through political or other action can be expected to be strategic. The theory of second-best allocations has been proposed in this context. Redistribution of endowment is excluded, and the mechanism is restricted to be a price mechanism, but the price system faced by consumers is allowed to be different from that faced by producers; all agents behave according to the rules of the (static) competitive mechanism. The allocations that satisfy these conditions, when the price systems are variable, are maximal allocations in the sense that they are Pareto optimal within the restricted class just defined. These are so-called *second-best* allocations. This analysis was pioneered by Lipsey and Lancaster (1956) and Diamond and Mirrlees (1971).

Efficient Allocation in Non(Neo)Classical Environments

The term *nonclassical* refers to those environments that fail to have the properties of classical ones; there may be indivisible commodities, nonconvexities in consumption sets, preferences or production sets, or externalities in production or consumption. An example of nonconvex preference would arise if a consumer preferred living in either Los Angeles or New York to living half the time in each city, or living halfway between them, depending on the way the commodity involved is specified. A production set representing a process that affords increasing returns to scale is an example of nonconvexity in production. A large investment project such as a road system is an example of a significant indivisibility. Phenomena of air or water pollution provide many examples of externalities in consumption and production.

The characterization of optimal allocation in terms of prices provided by the classical welfare theorems does not extend to nonclassical environments. If there are indivisibilities, equilibrium prices may fail to exist. Lerner (1934, 1947) has proposed a way of optimally allocating resources in the presence of indivisibilities. It would typically require adding up consumers' and producers' surplus.

Increasing returns to scale in production generally results in non-existence of competitive equilibrium, because of unbounded profit when prices are treated as given. Nash equilibrium, a concept from the theory of games, can exist even in cases of increasing returns. The difficulty is that such equilibria need not be optimal. Similar difficulties occur in cases of externalities.

Failure of the competitive price mechanism to extend the properties summarized in the classical welfare theorems to nonclassical environments has led economists to look for alternative ways of achieving optimal allocation in such cases. Such attempts have for the most part sought institutional arrangements that can be shown to result in optimal allocation. Ledyard (1968, 1971) analysed a mechanism for achieving Pareto optimal performance in environments with externalities. The use of taxes and subsidies advocated by Pigou (1932) to achieve Pareto optimal outcomes in cases of externalities is such an example. In a similar spirit Davis and Whinston (1962) distinguish externalities in production that leave marginal costs unaffected from those that do change marginal costs. In the former case they propose a pricing scheme, but one that involves lump-sum transfers. Marginal cost pricing, including lump-sum transfers to compensate for losses, which was extensively discussed as a device to achieve optimal allocation in the presence of increasing returns (Lerner 1944; Hotelling 1938; and many others) is another example of a scheme to realize optimal outcomes in nonclassical environments in a way that seeks to capture the benefits associated with decentralized resource allocation. In the case of production under conditions of increasing returns, the use of nonlinear prices has been suggested in an effort to achieve optimality with at least some of the benefits of decentralization. (See Arrow and Hurwicz 1960; Heal 1971; Brown and Heal 1982; Brown et al. 1986; Jennergren 1971; Guesnerie 1975.)

In the case of indivisibilities, and in the context of productive efficiency, integer programming algorithms exist for finding optima in specific problems, but a general characterization in terms of prices such as exists for the classical environments is not available. A decentralized process,

involving the use of randomization, whose equilibria coincide with the set of Pareto optima has been put forward by Hurwicz et al. (1975). This process has the property that the counterparts of the classical welfare theorems hold for environments in which all commodities are indivisible, and the set of feasible allocations is finite, or in which there are no indivisible commodities, or externalities, but there may be nonconvexities in production or consumption sets, or in preferences. This, of course, includes the possibility of increasing returns to scale in production.

The schemes and processes that have been proposed, including many not described here, are quite different from one another. If attention is confined to pricing schemes without additional elements, such as lump-sum transfers, it may be satisfactory to proceed on the basis of an informal intuitive notion of decentralization. This amounts in effect to identifying decentralization with the competitive mechanism, or more generally with price or market mechanisms. If a broader class of processes is to be considered, including some already mentioned in this discussion, then a formal concept of decentralized resource allocation process is needed.

Efficient Allocation Through Informationally Decentralized Processes

A formal definition of a concept of *allocation process* was first given by Hurwicz (1960). He also gave a definition of *informational decentralization* applying to a broad class of allocation mechanisms, based in part on a discussion by Hayek (1945) of the advantages of the competitive market mechanism for communicating knowledge initially dispersed among economic agents so that it can be brought to bear on the decisions that determine the allocation of resources. Hurwicz's formulation is as follows.

There is an initial dispersion of information about the environment; each agent is assumed to observe directly his own characteristic, e^i, but to know nothing directly about the characteristics of any other agent. In the absence of externalities, specifying the array of individual characteristics specifies the environment, i.e. $e = (e^1, \ldots, e^n)$. When there are externalities, an array of individual characteristics, each component of which corresponds to a possible environment, may not together constitute a possible environment. In more technical language, when there are externalities the set of environments is not the Cartesian product of its projections onto the sets of individual characteristics.

The goal of economic activity, whether efficiency, Pareto optimality or some other desideratum such as fairness, can be represented by a relation between the set of environments and the set of allocations, or outcomes. This relation assigns to each environment the set of allocations that meet the criterion of desirability. In the case of the Pareto criterion, the set of allocations that are Pareto optimal in a given environment is assigned to that environment. Formally, this relation is a correspondence (a set-valued function) from the set of environments to the set of allocations.

An allocation process, or mechanism, is modelled as an explicitly dynamic process of communication, leading to the determination of an outcome. In formal organizations standardized forms are frequently used for communication; in organized markets like the Stock Exchange, these include such things as order forms; in a business, forms on which weekly sales are reported; in the case of the Internal Revenue Service, income tax forms. A form consists of entries or blanks to be filled in a specified way. Thus, a form can be regarded as an ordered array of variables whose values come from specified sets. In the Hurwicz model, each agent is assumed to have a *language*, denoted M^i for the ith agent, from which his (possibly multi-dimensional) *message*, m^i, is chosen. The *joint message* of all the agents, $m = (m^1, \ldots, m^n)$ is in the *message space* $M = M^1 \times \ldots \times M^n$. Communication takes place in time, which is discrete; the message $m_t = \left(m_t^1, \ldots, m_t^n\right)$ denotes the message at time t. The message an agent emits at time t can depend on anything he knows at that time. This consists of what the agent knows about the environment by direct observation, by assumption, (*privacy*) his own characteristics, e^i for agent i, and what he has learned from others via the messages received from them. The agents' behaviour

is represented by *response functions*, which show how the current message depends on the information at hand. Agent i's message at time t is

$$m_i^i = f^i\left(m_{t-1}, m_{t-2}, \ldots, e^i\right), i = 1, \ldots, n, \quad t = 0, \ 1, \ 2, \ldots$$

If it is assumed that memory is finite, and bounded, it is possible without loss of generality to take the number of past periods remembered to be one. (If memory is unbounded, taking the number of periods remembered to be one excludes the possibility of a finite dimensional message space.) In that case the response equations become a system of first order temporally homogeneous difference equations in the messages. Thus:

$$m_i^i = \ f_i\left(m_{t-1}; e^i\right) \ \ i = 1, \ldots, n, t = 0, \ldots,$$

which can be written more compactly as

$$(^*) m_t = f(m_{t-1}; e).$$

(This formulation can accommodate the case of directed communication, in which some agents do not receive some messages; if agent i is not to receive the message of j, then f^i is independent of m^j, although m^j appears formally as an argument.) Analysis of informational properties of mechanisms is to begin with separated from that of incentives. When the focus is on communication and complexity qsts, the response functions are not regarded as chosen by the agent, but rather by the designer of the mechanism.

The iterative interchange of messages modelled by the difference equation system (*) eventually comes to an end, by converging to a stationary message. (It is also possible to have some stopping rule, such as to stop after a specified number of iterations.) The stationary message, which will be referred to as an *equilibrium message*, is then translated into an outcome, by means of the *outcome function*:

$$h : M \to Z,$$

where Z is the space of outcomes, usually allocations or trades. An allocation mechanism so modelled is called an *adjustment process*; it consists of the triple (M, f, h). Since no production or consumption takes place until all communication is completed, these processes are *tâtonnement* processes.

A more compact and general formulation was given by Mount and Reiter (1974) by looking only at message equilibria when attention is restricted to static properties. A correspondence is defined, called the *equilibrium message correspondence*. It associates to each environment the set of equilibrium messages for that environment. In order to satisfy the requirement of privacy, namely that each gent's message depend on the environment only through the agent's characteristic, the equilibrium message correspondence must be the intersection of individual message correspondences, each associating a set of message acceptable to the individual agent as equilibria in the light of his own characteristic. Thus the equilibrium message correspondence

$$\mu : E \to M,$$

is given by

$$\mu(e) = \bigcap_i \mu^i(e^t),$$

where $\mu^i : E^i \to M$, is the individual message correspondence of agent i. Note that here the message space M need not be the Cartesian product of individual languages. In the case of an adjustment process, the equilibrium message correspondence is defined by the conditions

$$\mu^i\left(e^i\right) = \left\{m \ \ in \ \ M | f^i\left(m; e^i\right) = m^i\right\}, \quad i = 1, \ldots, n$$

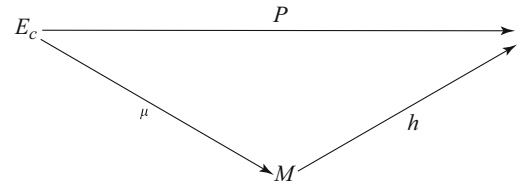
together with the condition that μ is the intersection of the μ^i. Specification of the outcome function $h : M \to Z$ completes the model, (M, μ, h).

The performance of a mechanism of this kind can be characterized by the mapping defined by the composition of the equilibrium message correspondence μ and the outcome function h. The mapping $h\mu; E \to Z$, possibly a correspondence, specifies the outcomes that the mechanism (M, μ, h) generates in each environment in E. A mechanism, whether in

the form of an adjustment process, or in the equilibrium form, is called *Pareto-satisfactory* (Hurwicz 1960) if for each environment in the class under consideration, the set of outcomes generated by the mechanism coincides with the set of Pareto optimal outcomes for that environment. Allowance must be made for redistribution of initial endowment, as in the case of the second welfare theorem. (A formulation in the framework of mechanisms is given in Mount and Reiter 1977.)

The competitive mechanism formalized as a static mechanism is as follows. (Hurwicz 1960, has given a different formulation, and Sonnenschein 1974, has given an axiomatic characterization of the competitive mechanism from a somewhat different point of view.) The message space M is the space of prices and quantities of commodities going to each agent (it has dimension $n(l-1)$ when there are n agents and l commodities, taking account of budget constraints and Walras' Law), the individual message correspondence μ^i maps agent i's characteristic e^i to the graph of his excess demand function. The equilibrium message is the intersection of the individual ones, and is therefore the price-quantity combinations that solve the system of excess demand equations. The outcome function h is the projection of the equilibrium message onto the quantity components of M. Thus $h\mu(e)$ is a competitive equilibrium allocation (or trade) when the environment is e. The classical welfare theorems state that for each e in E_c, $h[\mu(e)] = p(e)$, where E_c denotes the set of classical environments and P is the Pareto correspondence. (Allowance must be made for redistribution of initial endowment in connection with the second welfare theorem. Explicit treatment of this is omitted to avoid notational complexity. The decentralized redistribution of initial endowment is, as in the case of the second welfare theorem, not addressed.) The welfare theorems can be summarized in the Mount-Reiter diagram (Fig. 2) (Reiter 1977).

The welfare theorems state that this diagram *commutes* in the sense that starting from any environment e in E_c one reaches the same allocations via the mechanism, that is, via $h\mu$, as via the Pareto correspondence P.



Efficient Allocation, Fig. 2

With welfare theorems as a guide, the class of environments E_c can be replaced by some other class E, and the Pareto correspondence can be replaced by a correspondence, P, embodying another criterion of optimality, and one can ask whether there is a mechanism, (M, μ, h) that makes the diagram commute, or, in other words, *realizes* P? Without further restrictions on the mechanism, this is a triviality, because one agent can act as a central agent to whom all others communicate their environmental characteristics; the central agent then has the information required to evaluate P.

The concept of an *informationally decentralized mechanism* defined by Hurwicz (1960) makes explicit intuitive notions underlying the view that the price mechanism is decentralized.

Informationally decentralized processes are a subclass of so-called *concrete processes*, introduced by Hurwicz (1960). These are processes that use a language and response rules that allow production and distribution plans to be specified explicitly. The informationally decentralized processes are those whose response rules permit agents to transmit information only about their own actions, and which in effect require each agent to treat the rest of the economy either as one aggregate, or in a symmetrical way that, like the aggregate, gives anonymity to the other agents.

In the case of static mechanisms, the requirements for informational decentralization boil down to the condition that the message space have no more than a certain finite dimension, and in some cases only that it be of finite dimension. In the case of classical environments this can be seen to include the competitive mechanism, and to exclude the obviously centralized one mentioned above.

Without going deeply into the matter, an objective of this line of research is to analyse explicitly

the consequences of constraints on economic organization that come from limitations on the capacity of economic agents to observe, communicate and process information. One important result in this field is that there is no mechanism (M, μ, h) where μ preserves privacy, that uses messages smaller (in dimension) than those of the competitive mechanism (Hurwicz 1972b; Mount and Reiter 1974; Walker 1977; Osana 1978). Similar results have been obtained for environments with public goods, showing that the Lindahl mechanism uses the minimal message space (Sato 1981). Another objective is to analyse effects on incentives arising from private motivations in the presence of private information; that is, information held by one agent that is not observable by others, except perhaps at a cost. (There is a large literature on this subject under the rubric 'incentive compatibility', or 'strategic implementation' (Dasgupta et al. 1979; Hurwicz 1971, 1972a). The informational requirements of achieving a specified performance taking some aspects of incentive compatibility into account have been studied by Hurwicz (1976), Reichelstein (1984a, 1984b) and by Reichelstein and Reiter (1985).

Some important results for non-neoclassical environments can be mentioned. Hurwicz (1960, 1972a) has shown that there can be no informationally decentralized mechanism that realizes Pareto optimal performance on a class of environments that includes those with externalities. Calsamiglia (1977, 1982) has shown in a model of production that if the set of environments includes a sufficiently rich class of those with increasing returns to scale in production, then the dimension of the message space of any mechanism that realizes efficient production cannot be bounded.

Efficient Allocation with Infinitely Many Commodities

An infinite dimensional commodity space is needed when it is necessary to make infinitely many distinctions among goods and services. This is the case when commodities are distinguished according to time of availability and the time horizon in the model is not bounded or when time is continuous, or according to location when there is more than a finite number of possible locations; differentiated commodities provide other examples, and so does the case of uncertainty with infinitely many states. The bulk of the literature deals with the infinite horizon model of allocation over time, though recently more attention is given to models of product differentiation. Ramsey (1928) studied the problem of saving in a continuous time infinite horizon model with one consumption good and an infinitely lived consumer. He used as the criterion of optimality the infinite sum (integral) of undiscounted utility. Ramsey's contribution was largely ignored, and rediscovered when attention returned to problems of economic growth. A model of maximal sustainable growth based on a linear technology with no unproduced inputs was formulated by von Neumann (1937 in German; English translation, 1945–6). This contribution was unknown among English-speaking economists until after World War II. Study of intertemporal allocation by Anglo-American economists effectively began with the contributions of Harrod (1939) and Domar (1946). These models were concerned with stationary growth at a constant sustainable rate (stationary growth paths) rather than full intertemporal efficiency. Malinvaud (1953) first addressed this problem in a pioneering model of intertemporal allocation with an infinite horizon.

Efficient allocation over (discrete) time would be covered by the finite dimensional models described above if the time horizon were finite. It might be thought that a model with a sufficiently large but still finite horizon would for all practical purposes be equivalent to one with an infinite horizon, while avoiding the difficulties of infinity, but this is not the case, because of the dependence of efficient or optimal allocations on the value given to final stocks, a value that must depend on their uses beyond the horizon.

Malinvaud (1953) formulated an important infinite horizon model, which is the infinite dimensional counterpart of the linear activity analysis model of Koopmans. In Malinvaud's model time is discrete. The time horizon consists of an infinite sequence of time periods. At each

date there are finitely many commodities. All commodities are desired in each time period, and no distinction is made between desired, intermediate and primary commodities. As in the activity analysis model, there is no explicit reference to preferences of consumers. Productive efficiency over time is analysed in terms of the output available for consumption, rather than the resulting utility levels.

Technology is represented by a production set X^t for each time period $t = 1, 2, \ldots$, an element of X^t being an ordered pair (a^t, b^{t+1}) of commodity bundles where a^t represents inputs to a production process in period t, and b^{t+1} represents the outputs of that process available at the beginning of period $t+1$. Here both a^t and b^{t+1} are non-negative. The set X^t is the aggregate production set for the economy during period t. The net outputs available for consumption are given by

$$y^t = b^t - a^t, \text{for} \quad t \geq 1,$$

where b^1 is the initial endowment of resources available at the beginning of period 1. A *programme* is an infinite sequence $\langle (a^t, b^{t+1}) \rangle$; it is a *feasible programme* if (a^t, b^{t+1}); is in X^t, and $b^t - a^t \geq 0$ for each $t \geq 1$, given b^1. The sequence $y = \langle y^t \rangle$ is called the *net output programme* associated with the given programme; it is a *feasible net output programme* if it is the net output programme of a feasible programme. A programme is *efficient* if it is (1) feasible and (2) there is no other programme that is feasible, from the same initial resources b^1, and provides at least as much net output in every period and a larger net output in some period. This is the concept of efficient production, already seen in the linear activity analysis model, now extended to an infinite horizon model. The main aim of this research is to extend to the infinite horizon model the characterization of efficient production by prices seen in the finite model. This goal is not quite reached, as is seen in what follows.

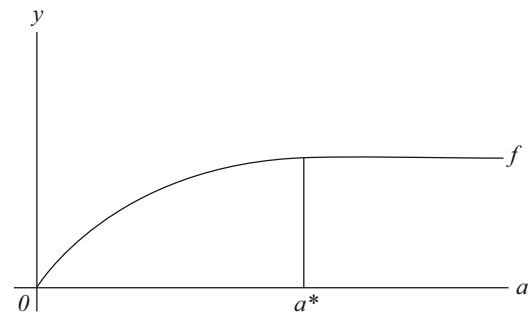
The main difficulties presented by the infinite horizon are already present in a special case of the Malinvaud model with one good and no consumers. Let Y be the set of all non-negative sequences $y = (y_t)$ that satisfy $0 \leq y_t = f(a_{t-1}) - a_t$ for $t \geq 1$, and $0 \leq y^0 = b^1 - a^0$, $b^1 > 0$, where f is a real-valued continuous concave function on the non-negative real numbers (the production function), $f(0) = 0$, and b^1 is the given initial stock. The set Y is the set of all feasible programmes. A programme $y' - y > 0$. A price system is an infinite sequence $p = (p^t)$ of non-negative numbers. Denote by P the set of all price systems.

Malinvalud recognized the possibility that an efficient net output programme (y^t) need not have an associated system of non-zero prices (p^t) relative to which the production programme generating y satisfies the condition of intertemporal profit maximization, namely that

$$P^{t+1} f(a^t) - p^t a^t \geq p^{t+1} f(a) - p^t a$$

for all t and every $a \geq 0$. (Here (a^t) is the sequence of inputs producing y.) A condition introduced by Malinvaud, called *nontightness*, is sufficient for the existence of such non-zero prices. Alternative proofs of Malinvaud's existence theorem were given by Radner (1967) and Peleg and Yaari (1970). (An example showing the possibility of non-existence given by Peleg and Yaari (1970) is as follows. Suppose f is as shown in Fig. 3.

At an interior efficient, and therefore value maximizing, programme the first-order necessary conditions for a maximum imply $p^{t+1} \dot{f}(a^t) = p^t$. If there is a time at which $a^t = a^*$, in an efficient programme, then, since $\dot{f}(a^*) = 0$ it follows that prices at all prior and future times are 0. (Non-tightness rules out such examples.)



Efficient Allocation, Fig. 3

On the side of sufficiency, Malinvaud showed that intertemporal profit maximization relative to a strictly positive price system p is not enough to ensure that a feasible programme is efficient. An additional (transversality) condition is needed. In the present model the following to such a condition;

$$\lim_{t\to\infty} p^t y^t = 0.$$

Cass (1972) has given a criterion that completely characterizes the set of efficient programmes in a one-good model with strictly concave and smooth production technology that satisfies endpoint conditions $0 \leq f'(\infty) < 1 < f'(x) < \infty$ for some $x > 0$. Cass's criterion, states that a programme is *inefficient* if and only if the associated competitive prices – that is, satisfying $p^{t+1} f'(a^t) = p^t$ – also satisfy $\Sigma_{t=1}^{\infty}(1/p^t) < \infty$. This criterion may be interpreted as requiring the terms of trade between present and future to deteriorate sufficiently fast. Other similar conditions have been presented (Benveniste and Gale 1975; Benveniste 1976; Majumdar 1974; Mitra 1979). It is hard to see how any transversality condition can be interpreted in terms of decentralized resource allocation.

An alternate approach to characterizing efficient programmes was taken by Radner (1967), based on value functions as introduced in connection with valuation equilibrium by Debreu (1954). (Valuation equilibrium was discussed in connection with Arrow's exceptional case, above.) The value function approach was followed up by Majumdar (1970, 1972) and by Peleg and Yaari (1970). A price system defines a continuous linear functional, (a real-valued linear function) on the commodity space. This function assigns to a programme its present value. The present value may not be well-defined, because the infinite sequence that gives it diverges. This creates certain technical problems passed over here. A more important difficulty is that linear functionals exist that are not defined by price systems. Radner's approach was to characterize efficient programmes in terms of maximization of present value relative to a linear functional on the commodity space. Radner showed, technical matters aside, that:

1. If a feasible programme maximizes the value of net output (consumption) relative to a strictly positive continuous linear functional, then it is efficient.
2. If a given programme is efficient, then there is a nonzero non-negative continuous linear functional such that the given programme maximizes the value of net output relative to that functional on the set of feasible programmes.

These propositions seem to be the precise counterparts of the ones characterizing efficiency in the finite horizon model. Unfortunately, a linear functional may not have a representation in the form of the inner product of a price sequence with a net output sequence. (The production function $f(a) = a^{\beta}$, with $0 < \beta < 1$ provides an example. It is known that the programme with constant input sequence $x_t = (1/\beta)^{\beta/\beta - 1}$ and output sequence $y_t = (1/\beta)^{\beta/\beta - 1} - (1/\beta)^{1/\beta - 1} t = 1, 2, \ldots,$ is efficient, and therefore there is a continuous linear functional relative to which it is value maximizing. But there is no price sequence (p^t) that represents that linear functional.) This presents a serious problem, because in the absence of such a representation it is unclear whether this characterization has an interpretation in terms of decentralized allocation processes; profit in any one period can depend on 'prices at infinity'.

This approach has the advantage that it is applicable not only to infinite horizon models, but to a broader class in which the commodity space is infinite dimensional. Bewley (1972), Mas-Colell (1977) and Jones (1984) among others discuss Pareto optimality and competitive equilibrium in economies with infinitely many commodities. Hurwicz (1958) and others analysed optimal allocation in terms of nonlinear programming in infinite dimensional spaces. Theorems of programming in infinite dimensional spaces are also used in some of the models mentioned in this discussion.

The basic difficulties encountered in the one-good model, apart from the numerous technical problems that tend to make the literature large and

diverse as different technical structures are investigated, are on the one hand the fact that transversality conditions are indispensable, and on the other the possibility that linear functionals, even when they exist, may not be representable in terms of price sequences. These problems raise strong doubt about the possibility of achieving efficient intertemporal resource allocation by decentralized means, though they leave open the possibility that some other decentralized mechanism, not using prices, might work. Analysis of this possibility has just begun, and is discussed below.

The difficulties seen in the one-good production model persist in more elaborate ones, including multisectoral models with efficiency as the criterion, and models with consumers in which Pareto optimality is the criterion. McFadden et al. (1980) studied a model in which there are firms, and overlapping generations of consumers, as in the model first investigated by Samuelson (1958). Each consumer lives for a finite time and has a consumption set and preferences like the consumers in a finite horizon model. A model with overlapping generations of consumers presents the fundamental difficulty that consumers cannot trade with future consumers as yet unborn. This difficulty can appear even in a finite horizon model if there are too few markets. The economy is closed in the sense that there are no non-produced resources; the von Neumann growth model is an example of such a model. Building on the results of an earlier investigation (Majumdar et al. 1976), these authors introduced several notions of price systems, of competitive equilibrium, efficiency and optimality, and sought to establish counterparts of the classical welfare theorems. To summarize, in the 1976 paper they strengthen an earlier result of Bose (1974) to the effect that the problem of proper distribution of goods in essentially a short-run problem, and that the only long-run problem, one created by the infinite horizon, is that of inefficiency through overaccumulation of capital. In the 1980 paper the focus is on the relationships among various notions of equilibrium and Pareto optimality. The force of their results is, as might be expected, that the difficulties already seen in one-good model without consumers persist in this model. A transversality condition is made part of the definition of competitive equilibrium in order to obtain the result that an equilibrium is optimal. A partial converse requires some additional assumptions on the technology (reachability) and on the way the economy fits together (nondecomposability). These results certainly illuminate the infinite horizon model with overlapping generations of consumers and producers, but the possibility of efficient or optimal resource allocation by decentralized means is not different from that in the one-good Malinvaud model.

Hurwicz and Majumdar in an unpublished manuscript dated 1983, and later Hurwicz and Weinberger (1984), have addressed this issue directly, building on the approach of mechanism theory.

Hurwicz and Majumdar have studied the problem of efficiency in a model with an infinite number of periods. In each period there are finitely many commodities, one producer who is alive for just one period, and no consumers' choices. The criterion is the maximization of the discounted value of the programme (well-defined in this model). The producer alive in any period knows only the technology in that period. The question is whether there is a (static) privacy preserving mechanism using a finite dimensional message space whose equilibria coincide with the set of efficient programmes. The question can be put as follows. In each period a message is posted. The producer alive in that period responds 'Yes' or 'No'. If every producer over the entire infinite horizon answers 'Yes', the programme is an outcome corresponding to the equilibrium consisting of the infinite succession of posted messages. Since each producer knows only the technology prevailing in the period when he is alive, the process preserves privacy. If in addition the message posted in each period is finite dimensional, the process is informationally decentralized. Period-by-period profit maximization using period-by-period prices is a mechanism of this type; the message posted in each period consists of the vector of prices for that period, and the production plan for that period, both finite dimensional. The object is to characterize all efficient programmes as equilibria of such a mechanism.

This would be an analogue of the classical welfare theorems, but without the restriction to mechanisms that use prices in their messages.

The main result is in the nature of an impossibility theorem. If the technology is constant over time, and that fact is common knowledge at the beginning, the problem is trivial since knowledge of the technology in the first period automatically means knowledge of it in every period. On the other hand, if there is some period whose technology is not known in the first period, then there is no finite dimensional message that can characterize efficient programmes, and in that sense, production cannot be satisfactorily decentralized over time.

Hurwicz and Weinberger (1984) have studied a model with both producers and consumers. As with producers, there is a consumer in each period, who lives for one period. The consumer in each period has a one-period utility function, which is not known by the producer; similarly the consumer does not know the production function. The criterion of optimality is the maximization of the sum of discounted utilities over the infinite horizon. Hurwicz and Weinberger show that there is no privacy preserving mechanism of the type just described whose equilibria correspond to the set of optimal programmes. It should be noted that their mechanism requires that the first-period actions (production, consumption and investment decisions) be made in the first period, and not be subject to revision after the infinite process of verification is completed. (On the other hand, under tâtonnement assumptions it may be possible to decentralize. In this model t*â*tonnement entails reconsideration 'at infinity'.)

If attention is widened to efficient programmes, and if technology is constant over time, there is an efficient programme with a fixed ratio of consumption to investment. This programme can be obtained as the equilibrium outcome of a mechanism of the specified type. However, this corresponds to only one side of the classical welfare theorems. It says that the outcome of such a mechanism is efficient; but it does not ensure that every efficient programme can be realized as the outcome of such a mechanism. The latter property fails in this model.

See Also

- ▶ Incentive Compatibility
- ▶ Linear Programming
- ▶ Welfare Economics

Bibliography

Arrow, K. 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., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22 (July): 265–290.

Arrow, K., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Arrow, K., and L. Hurwicz. 1960. Decentralization and computation in resource allocation. In *Essays in economics and econometrics*, ed. R.W. Pfouts, 34–104. Chapel Hill: University of North Carolina Press.

Arrow, K., L. Hurwicz, and H. Uzawa. 1961. Constraint qualifications in maximization problems. *Naval Research Logistics Quarterly* 8 (2): 175–191.

Benveniste, L. 1976. Two notes on the Malinvaud condition for efficiency of infinite horizon programs. *Journal of Economic Theory* 12: 338–346.

Benveniste, L., and D. Gale. 1975. An extension of Cass' characterization of infinite efficient production programs. *Journal of Economic Theory* 10: 229–238.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4: 514–540.

Bose, A. 1974. *Pareto optimality and efficient capital accumulation*. Discussion paper no. 74–4. Department of Economics, University of Rochester.

Brown, D., and G. Heal 1982. Existence, local-uniqueness and optimality of a marginal cost pricing equilibrium in an economy with increasing returns. *Cal. Tech*. Social Science working paper no. 415.

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: 371–379.

Calsamiglia, X. 1977. Decentralized resource allocation and increasing returns. *Journal of Economic Theory* 14: 263–283.

Calsamiglia, X. 1982. On the size of the message space under non-convexities. *Journal of Mathematical Economics* 10: 197–203.

Cass, D. 1972. On capital over-accumulation in the aggregative neoclassical model of economic growth: A complete characterization. *Journal of Economic Theory* 4 (2): 200–223.

Dantzig, G.B. 1951a. The programming of interdependent activities. In *Activity analysis of production and*

allocation, ed. T. Koopmans, 19–32, Cowles Commission Monograph No. 13. New York: Wiley, ch. 2.

Dantzig, G.B. 1951b. Maximization of a linear function of variables subject to linear inequalities. In *Activity analysis of production and allocation*, ed. T. Koopmans, 339–347, Cowles Commission Monograph No. 13, New York: Wiley, ch. 21.

Dasgupta, P., P. Hammond, and E. Maskin. 1979. The implementation of social choice rules: Some general results on incentive compatibility. *Review of Economic Studies* 46: 185–216.

Davis, O.A., and A.B. Whinston. 1962. Externalities welfare and the theory of games. *Journal of Political Economy* 70: 214–262.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences of the USA* 40 (7): 588–592.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38 (3): 387–392.

Diamond, P., and J. Mirrlees. 1971. Optimal taxation and public production. I: Production efficiency; II: Tax rules. *American Economic Review* 61: 8–27. 261–78.

Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14 (April): 137–147.

Fenchel, W. 1950. Convex cones, sets, and functions. Princeton University (hectographed).

Guesnerie, R. 1975. Pareto optimality in non-convex economies. *Econometrica* 43: 1–29.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Heal, G. 1971. Planning, prices and increasing returns. *Review of Economic Studies* 38: 281–294.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–269.

Hurwicz, L. 1958. Programming in linear spaces. In *Studies in linear and non-linear programming*, ed. K. Arrow, L. Hurwicz, and H. Uzawa. Stanford: Stanford University Press.

Hurwicz, L. 1960. Optimality and informational efficiency in resource allocation processes. In *Mathematical methods in the social sciences, 1959*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Hurwicz, L. 1971. Centralization and decentralization in economic processes. In *Comparison of economic systems: Theoretical and methodological approaches*, ed. A. Eckstein. Berkeley: University of California Press. ch. 3.

Hurwicz, L. 1972a. On informationally decentralized systems. In *Decision and organization*, ed. C. McGuire and R. Radner, 297–336. Amsterdam/London: North-Holland, ch. 14.

Hurwicz, L. 1972b. On the dimensional requirements of in formationally decentralized Pareto-satisfactory processes. Presented at the conference seminar in decentralization North-western University. In *Studies in resource allocation processes*, ed. K.J. Arrow and L. Hurwicz. Cambridge: Cambridge University Press. 1977.

Hurwicz, L. 1976. On informational requirements for nonwasteful resource allocation systems. In *Mathematical models in economics: Papers and proceedings of a US-USSR seminar, Moscow*, ed. S. Shulman. New York: National Bureau of Economic Research.

Hurwicz, L. and H. Weinberger 1984. Paper presented at IMA seminar in Minneapolis.

Hurwicz, L., Radner, R., and Reiter, S. 1975. A stochastic decentralized resource allocation process. *Econometrica* 43: Part I, 187–221; Part II, 363–93.

Jennergren, L. 1971. *Studies in the mathematical theory of decentralized resource-allocation*. PhD dissertation, Stanford University.

Jones, L. 1984. A competitive model of commodity differentiation. *Econometrica* 52: 507–530.

Kantorovitch, L. 1939. Matematicheskie metody organizatii i planirovania proizvodstva (Mathematical methods in the organization and planning of production). Izdanie Leningradskogo Gosudarstvennogo Universiteta, Leningrad. Trans. in *Management Science* 6(4), July 1960, 363–422.

Kantorovitch, L. 1942. On the translocation of masses (In English.) *Comptes Rendus (Doklady) de l'Academie des Sciences d l'URSS* 37(7–8).

Kantorovitch, L., and M. Gavurin 1949. Primenenie matematicheskikh metodov v voprosakh analyza grusopotokov (The application of mathematical methods to problems of freight flow analysis). In *Problemy Povysheniia Effektivnosty Raboty Transporta* (Problems of raising the efficiency of transportation), ed. V. Zvonkov. Moscow/Leningrad: Izdatel'stvo Akademii Nauk SSSR.

Koopmans, T.C. 1951a. Analysis of production as an efficient combination of activities. In *Activity analysis of production and allocation*, ed. T. Koopmans, 33–37, Cowles Commission Monograph No. 13, New York: Wiley, ch. 3.

Koopmans, T.C. 1951b. Efficient allocation of resources. *Econometrica* 19: 455–465.

Koopmans, T.C. 1957. *Three essays on the state of economic science*, 66–104. New York: McGraw-Hill.

Kuhn, H., and A. Tucker. 1951. Nonlinear programming. In *Proceedings of the second Berkeley Symposium on mathematical statistics and probability*, ed. J. Neyman, 481–492. Berkeley: University of California Press.

Ledyard, J. 1968. Resource allocation in unselfish environments. *American Economic Review* 58: 227–237.

Ledyard, J. 1971. A convergent Pareto-satisfactory non--tâtonnement adjustment process for a class of unselfish exchange environments. *Econometrica* 39: 467–499.

Lerner, A. 1934. The concept of monopoly and measurement of monopoly power. *Review of Economic Studies* 1 (3): 157–175.

Lerner, A. 1944. *The economics of control*. New York: Macmillan.

Lipsey, R., and K. Lancaster. 1956. The general theory of second best. *Review of Economic Studies* 24: 11–32.

Majumdar, M. 1970. Some approximation theorems on efficiency prices for infinite programs. *Journal of Economic Theory* 2: 399–410.
Majumdar, M. 1972. Some general theorems of efficiency prices with an infinite dimensional commodity space. *Journal of Economic Theory* 5: 1–13.
Majumdar, M. 1974. Efficient programs in infinite dimensional spaces: A complete characterization. *Journal of Economic Theory* 7: 355–369.
Majumdar, M., T. Mitra, and D. McFadden. 1976. On efficiency and Pareto optimality of competitive programs in closed multisector models. *Journal of Economic Theory* 13: 26–46.
Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 233–268.
Mas-Colell, A. 1977. Regular nonconvex economies. *Econometrica* 45: 1387–1407.
McFadden, D., T. Mitra, and M. Majumdar. 1980. Pareto optimality and competitive equilibrium in infinite horizon economies. *Journal of Mathematical Economics* 7: 1–26.
McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27 (1): 54–71.
Mitra, T. 1979. On optimal economic growth with variable discount rates: Existence and stability results. *International Economic Review* 20: 133–145.
Mount, K., and S. Reiter. 1974. The informational size of message spaces. *Journal of Economic Theory* 8: 161–192.
Mount, K., and S. Reiter. 1977. Economic environments for which there are Pareto satisfactory mechanisms. *Econometrica* 45: 821–842.
Neumann, J von. 1937. A model of general economic equilibrium. *Ergebnisse eines mathematischen Kolloquiums* (8). Trans. from German, *Review of Economic Studies* 13(1), (1945–6): 1–9.
Nikaido, H. 1969. *Convex structures and economic theory.* New York: Academic Press.
Nikaido, H. 1970. *Introduction to sets and mappings in modern economics*. Trans. K. Sato, Amsterdam: North-Holland (Japanese original, Tokyo, 1960).
Osana, H. 1978. On the informational size of message spaces for resource allocation processes. *Journal of Economic Theory* 17: 66–78.
Peleg, B., and M. Yaari. 1970. Efficiency prices in an infinite dimensional commodity space. *Journal of Economic Theory* 2: 41–85.
Pigou, A. 1932. *The economics of welfare*. 4th ed. London: Macmillan.
Radner, R. 1967. Efficiency prices for infinite horizon production programs. *Review of Economic Studies* 34: 51–66.
Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Reichelstein, S. 1984a. Dominant strategy implementation, incentive compatibility and informational requirements. *Journal of Economic Theory* 34 (1): 32–51.
Reichelstein, S. 1984b. *Information and incentives in economic organizations*. PhD dissertation, Northwestern University.
Reichelstein, S. and Reiter, S. 1985. *Game forms with minimal strategy spaces*. Discussion paper no. 663, The Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston, Ill.
Reiter, S. 1977. Information and performance in the (new)2 welfare economics. *American Economic Review* 67: 226–234.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P. 1949. *Market mechanisms and maximization, I, II, III,* Hectographed memoranda. Santa Monica: The RAND Corporation.
Samuelson, P. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66 (December): 467–482.
Sato, F. 1981. On the informational size of message spaces for resource allocation processes in economies with public goods. *Journal of Economic Theory* 24: 48–69.
Slater, M. 1950. Lagrange multipliers revisited: A contribution to non-linear programming. *Cowles Commission Discussion Paper*, Math. 403, also RM–676, 1951.
Sonnenschein, H. 1974. An axiomatic characterization of the price mechanism. *Econometrica* 42: 425–434.
Uzawa, H. 1958. The Kuhn-Tucker Theorem in concave programming. In *Studies in linear and non-linear programming*, ed. K. Arrow, L. Hurwicz, and H. Uzawa. Stanford: Stanford University Press.
von Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35, 519–553. Reprinted in F. von Hayek, *Individualism and economic order*, 77–92. Chicago: University of Chicago Press, 1949.
Walker, M. 1977. On the informational size of message spaces. *Journal of Economic Theory* 15: 366–375.

Efficient Markets Hypothesis

Andrew W. Lo

Abstract

The efficient markets hypothesis (EMH) maintains that market prices fully reflect all available information. Developed independently by Paul A. Samuelson and Eugene F. Fama in the 1960s, this idea has been applied extensively to theoretical models and empirical studies of financial securities prices, generating considerable controversy as well as fundamental insights into the price-discovery process. The

most enduring critique comes from psychologists and behavioural economists who argue that the EMH is based on counterfactual assumptions regarding human behaviour, that is, rationality. Recent advances in evolutionary psychology and the cognitive neurosciences may be able to reconcile the EMH with behavioural anomalies.

Keywords

Adaptive markets hypothesis; Agent-based models; Altruism; Arbitrage; Asset price anomalies; Behavioural biases; Behavioural economics; Behavioural finance; Bid–ask bounce; Biology and economics; Bounded rationality; Capital asset pricing model; Consumer choice; Creative destruction; Deductive inference; Dividend smoothing; Dividend-discount model; Dutch books; Economic complexity; Efficient markets hypothesis; Emotions; Equilibrium; Equity risk premium; Evolutionary economics; Evolutionary game theory; Evolutionary psychology; Fama, E. F.; Financial economics; Herding; Hyperbolic discounting; Inductive inference; Information aggregation; Informational efficiency; January effect; Joint hypotheses; Learning; Long-term memory; Loss aversion; Market efficiency; Martingales; Miscalibration of probabilities; Natural selection; Neuroeconomics; Noise traders; Optimization; Overconfidence; Overreaction; Post-earnings announcement drift; Preferences; Present value; Price reversals; Psychological accounting; Punctuated equilibrium; Random walk hypothesis; Rational expectations; Regret; Relative efficiency; Risk aversion; Risk preferences; Risk–reward relation; Samuelson, P. A.; Satisficing; Serial correlation; Simon, H.; Size effect; Social norms; Sociobiology; Statistical inference; Stock price volatility; Survival of the fittest; Uncertainty; Utility maximization; Variance bounds; Variance decomposition

JEL Classifications

G0

There is an old joke, widely told among economists, about an economist strolling down the street with a companion. They come upon a $100 bill lying on the ground, and as the companion reaches down to pick it up, the economist says, 'Don't bother – if it were a genuine $100 bill, someone would have already picked it up'. This humorous example of economic logic gone awry is a fairly accurate rendition of the efficient markets hypothesis (EMH), one of the most hotly contested propositions in all the social sciences. It is disarmingly simple to state, has far-reaching consequences for academic theories and business practice, and yet is surprisingly resilient to empirical proof or refutation. Even after several decades of research and literally thousands of published studies, economists have not yet reached a consensus about whether markets – particularly financial markets – are, in fact, efficient.

The origins of the EMH can be traced back to the work of two individuals in the 1960s: Eugene F. Fama and Paul A. Samuelson. Remarkably, they independently developed the same basic notion of market efficiency from two rather different research agendas. These differences would propel the them along two distinct trajectories leading to several other breakthroughs and milestones, all originating from their point of intersection, the EMH.

Like so many ideas of modern economics, the EMH was first given form by Paul Samuelson (1965), whose contribution is neatly summarized by the title of his article: 'Proof that Properly Anticipated Prices Fluctuate Randomly'. In an informationally efficient market, price changes must be unforecastable if they are properly anticipated, that is, if they fully incorporate the information and expectations of all market participants. Having developed a series of linear-programming solutions to spatial pricing models with no uncertainty, Samuelson came upon the idea of efficient markets through his interest in temporal pricing models of storable commodities that are harvested and subject to decay. Samuelson's abiding interest in the mechanics and kinematics of prices, with and without uncertainty, led him and his students to several fruitful research agendas including solutions for the dynamic asset-allocation and consumption-

savings problem, the fallacy of time diversification and log-optimal investment policies, warrant and option-pricing analysis and, ultimately, the Black and Scholes (1973) and Merton (1973) option-pricing models.

In contrast to Samuelson's path to the EMH, Fama's (1963, 1965a, b, 1970) seminal papers were based on his interest in measuring the statistical properties of stock prices, and in resolving the debate between technical analysis (the use of geometric patterns in price and volume charts to forecast future price movements of a security) and fundamental analysis (the use of accounting and economic data to determine a security's fair value). Among the first to employ modern digital computers to conduct empirical research in finance, and the first to use the term 'efficient markets' (Fama 1965b), Fama operationalized the EMH hypothesis – summarized compactly in the epigram 'prices fully reflect all available information' – by placing structure on various information sets available to market participants. Fama's fascination with empirical analysis led him and his students down a very different path from Samuelson's, yielding significant methodological and empirical contributions such as the event study, numerous econometric tests of single-and multi-factor linear asset-pricing models, and a host of empirical regularities and anomalies in stock, bond, currency and commodity markets.

The EMH's concept of informational efficiency has a Zen-like, counter-intuitive flavour to it: the more efficient the market, the more random the sequence of price changes generated by such a market, and the most efficient market of all is one in which price changes are completely random and unpredictable. This is not an accident of nature, but is in fact the direct result of many active market participants attempting to profit from their information. Driven by profit opportunities, an army of investors pounce on even the smallest informational advantages at their disposal, and in doing so they incorporate their information into market prices and quickly eliminate the profit opportunities that first motivated their trades. If this occurs instantaneously, which it must in an idealized world of 'frictionless' markets and costless trading, then prices must always fully reflect all available information. Therefore, no profits can be garnered from information-based trading because such profits must have already been captured (recall the $100 bill on the ground). In mathematical terms, prices follow martingales.

Such compelling motivation for randomness is unique among the social sciences and is reminiscent of the role that uncertainty plays in quantum mechanics. Just as Heisenberg's uncertainty principle places a limit on what we can know about an electron's position and momentum if quantum mechanics holds, this version of the EMH places a limit on what we can know about future price changes if the forces of economic self-interest hold.

A decade after Samuelson's (1965) and Fama's (1965a, b, 1970) landmark papers, many others extended their framework to allow for risk-averse investors, yielding a 'neoclassical' version of the EMH where price changes, properly weighted by aggregate marginal utilities, must be unforecastable (see, for example, Leroy 1973; Rubinstein 1976; Lucas 1978). In markets where, according to Lucas (1978), all investors have 'rational expectations', prices do fully reflect all available information and marginal-utility-weighted prices follow martingales. The EMH has been extended in many other directions, including the incorporation of non-traded assets such as human capital, state-dependent preferences, heterogeneous investors, asymmetric information, and transactions costs. But the general thrust is the same: individual investors form expectations rationally, markets aggregate information efficiently, and equilibrium prices incorporate all available information instantaneously.

The Random Walk Hypothesis

The importance of the EMH stems primarily from its sharp empirical implications many of which have been tested over the years. Much of the EMH literature before LeRoy (1973) and Lucas (1978) revolved around the random walk hypothesis (RWH) and the martingale model, two statistical descriptions of unforecastable price changes that were initially taken to be implications of the

EMH. One of the first tests of the RWH was developed by Cowles and Jones (1937), who compared the frequency of *sequences* and *reversals* in historical stock returns, where the former are pairs of consecutive returns with the same sign, and the latter are pairs of consecutive returns with opposite signs. Cootner (1962, 1964), Fama (1963, 1965a), Fama and Blume (1966), and Osborne (1959) perform related tests of the RWH and, with the exception of Cowles and Jones (who subsequently acknowledged an error in their analysis – Cowles 1960), all of these articles indicate support for the RWH using historical stock price data.

More recently, Lo and MacKinlay (1988) exploit the fact that return variances scale linearly under the RWH – the variance of a two-week return is twice the variance of a one-week return if the RWH holds – and construct a variance ratio test which rejects the RWH for weekly US stock returns indexes from 1962 to 1985. In particular, they find that variances grow faster than linearly as the holding period increases, implying positive serial correlation in weekly returns. Oddly enough, Lo and MacKinlay also show that individual stocks generally do satisfy the RWH, a fact that we shall return to below.

French and Roll (1986) document a related phenomenon: stock return variances over weekends and exchange holidays are considerably lower than return variances over the same number of days when markets are open. This difference suggests that the very act of trading creates volatility, which may well be a symptom of Black's (1986) noise traders.

For holding periods much longer than one week – for example, three to five years – Fama and French (1988) and Poterba and Summers (1988) find negative serial correlation in US stock returns indexes using data from 1926 to 1986. Although their estimates of serial correlation coefficients seem large in magnitude, there is insufficient data to reject the RWH at the usual levels of significance. Moreover, a number of statistical artifacts documented by Kim et al. (1991) and Richardson (1993) cast serious doubt on the reliability of these longer-horizon inferences.

Finally, Lo (1991) considers another aspect of stock market prices long thought to have been a departure from the RWH: long-term memory. Time series with longterm memory exhibit an unusually high degree of persistence, so that observations in the remote past are non-trivially correlated with observations in the distant future, even as the time span between the two observations increases. Nature's predilection towards long-term memory has been well-documented in the natural sciences such as hydrology, meteorology, and geophysics, and some have argued that economic time series must therefore also have this property.

However, using recently developed statistical techniques, Lo (1991) constructs a test for long-term memory that is robust to short-term correlations of the sort uncovered by Lo and MacKinlay (1988), and concludes that, despite earlier evidence to the contrary, there is little support for long-term memory in stock market prices. Departures from the RWH can be fully explained by conventional models of short-term dependence.

Variance Bounds Tests

Another set of empirical tests of the EMH starts with the observation that in a world without uncertainty the market price of a share of common stock must equal the present value of all future dividends, discounted at the appropriate cost of capital. In an uncertain world, one can generalize this *dividend-discount model* or *present-value relation* in the natural way: the market price equals the conditional expectation of the present value of all future dividends, discounted at the appropriate risk-adjusted cost of capital, and conditional on all available information. This generalization is explicitly developed by Grossman and Shiller (1981).

LeRoy and Porter (1981) and Shiller (1981) take this as their starting point in comparing the variance of stock market prices to the variance of *ex post* present values of future dividends. If the market price is the conditional expectation of present values, then the difference between the two, that is, the forecast error, must be uncorrelated with the conditional expectation by construction.

But this implies that the variance of the *ex post* present value is the sum of the variance of the market price (the conditional expectation) and the variance of the forecast error. Since volatilities are always non-negative, this variance decomposition implies that the variance of stock prices cannot exceed the variance of *ex post* present values. Using annual US stock market data from various sample periods, LeRoy and Porter (1981) and Shiller (1981) find that the variance bound is violated dramatically. Although LeRoy and Porter are more circumspect about the implications of such violations, Shiller concludes that stock market prices are too volatile and the EMH must be false.

These two papers ignited a flurry of responses which challenged Shiller's controversial conclusion on a number of fronts. For example, Flavin (1983), Kleidon (1986), and Marsh and Merton (1986) show that statistical inference is rather delicate for these variance bounds, and that, even if they hold in theory, for the kind of sample sizes Shiller uses and under plausible data-generating processes the sample variance bound is often violated purely due to sampling variation. These issues are well summarized in Gilles and LeRoy (1991) and Merton (1987).

More importantly, on purely theoretical grounds Marsh and Merton (1986) and Michener (1982) provide two explanations for violations of variance bounds that are perfectly consistent with the EMH. Marsh and Merton (1986) show that if managers smooth dividends – a well-known empirical phenomenon documented in several studies of dividend policy – and if earnings follow a geometric random walk, then the variance bound is violated in theory, in which case the empirical violations may be interpreted as *support* for this version of the EMH.

Alternatively, Michener constructs a simple dynamic equilibrium model along the lines of Lucas (1978) in which prices do fully reflect all available information at all times but where individuals are risk averse, and this risk aversion is enough to cause the variance bound to be violated in theory as well.

These findings highlight an important aspect of the EMH that had not been emphasized in earlier studies: tests of the EMH are always tests of joint hypotheses. In particular, the phrase 'prices fully reflect all available information' is a statement about two distinct aspects of prices: the information content and the price formation mechanism. Therefore, any test of this proposition must concern the *kind* of information reflected in prices, and *how* this information comes to be reflected in prices.

Apart from issues regarding statistical inference, the empirical violation of variance bounds may be interpreted in many ways. It may be a violation of EMH, or a sign that investors are risk averse, or a symptom of dividend smoothing. To choose among these alternatives, more evidence is required.

Overreaction and Underreaction

A common explanation for departures from the EMH is that investors do not always react in proper proportion to new information. For example, in some cases investors may overreact to performance, selling stocks that have experienced recent losses or buying stocks that have enjoyed recent gains. Such overreaction tends to push prices beyond their 'fair' or 'rational' market value, only to have rational investors take the other side of the trades and bring prices back in line eventually. An implication of this phenomenon is price reversals: what goes up must come down, and vice versa. Another implication is that *contrarian* investment strategies – strategies in which 'losers' are purchased and 'winners' are sold – will earn superior returns.

Both of these implications were tested and confirmed using recent US stock market data. For example, using monthly returns of New York Stock Exchange (NYSE) stocks from 1926 to 1982, DeBondt and Thaler (1985) document the fact that the winners and losers in one 36-month period tend to reverse their performance over the next 36-month period. Curiously, many of these reversals occur in January (see the discussion below on the 'January effect'). Chopra et al. (1992) reconfirm these findings after correcting for market risk and the size effect.

And Lehmann (1990) shows that a zero-net-investment strategy in which long positions in losers are financed by short positions in winners almost always yields positive returns for monthly NYSE/AMEX stock returns data from 1962 to 1985.

However, Chan (1988) argues that the profitability of contrarian investment strategies cannot be taken as conclusive evidence against the EMH because there is typically no accounting for risk in these profitability calculations (although Chopra et al. 1992 do provide risk adjustments, their focus was not on specific trading strategies). By risk-adjusting the returns of a contrarian trading strategy according to the capital asset pricing model, Chan (1988) shows that the expected returns are consistent with the EMH.

Moreover, Lo and MacKinlay (1990c) show that at least half of the profits reported by Lehmann (1990) are not due to overreaction but rather the result of positive cross-autocorrelations between stocks. For example, suppose the returns of two stocks A and *B* are both serially uncorrelated but are positively cross-autocorrelated. The lack of serial correlation implies no overreaction (which is characterized by negative serial correlation), but positive cross-autocorrelations yields positive expected returns to contrarian trading strategies. The existence of several economic rationales for positive cross-autocorrelation that are consistent with EMH suggests that the profitability of contrarian trading strategies is not sufficient evidence to conclude that investors overreact.

The reaction of market participants to information contained in earnings announcements also has implications for the EMH. In one of the earliest studies of the information content of earnings, Ball and Brown (1968) show that up to 80 per cent of the information contained in the earnings 'surprises' is anticipated by market prices.

However, the more recent article by Bernard and Thomas (1990) argues that investors sometimes underreact to information about future earnings contained in current earnings. This is related to the 'post-earnings announcement drift' puzzle first documented by Ball and Brown (1968), in which the information contained in earnings announcement takes several days to become fully impounded into market prices. Although such effects are indeed troubling for the EMH, their economic significance is often questionable – while they may violate the EMH in frictionless markets, very often even the smallest frictions – for example, positive trading costs, taxes – can eliminate the profits from trading strategies designed to exploit them.

Anomalies

Perhaps the most common challenge to the EMH is the anomaly, a regular pattern in an asset's returns which is reliable, widely known, and inexplicable. The fact that the pattern is regular and reliable implies a degree of predictability, and the fact that the regularity is widely known implies that many investors can take advantage of it.

For example, one of the most enduring anomalies is the 'size effect', the apparent excess expected returns that accrue to stocks of small-capitalization companies – in excess of their risks – which was first discovered by Banz (1981). Keim (1983), Roll (1983), and Rozeff and Kinney (1976) document a related anomaly: small capitalization stocks tend to outperform large capitalization stocks by a wide margin over the turn of the calendar year. This so-called 'January effect' seems robust to sample period, and is difficult to reconcile with the EMH because of its regularity and publicity. Other well-known anomalies include the Value Line enigma (Copeland and Mayers 1982), the profitability of short-term return-reversal strategies in US equities (Rosenberg et al. 1985; Chan 1988; Lehmann 1990; and Lo and MacKinlay 1990c), the profitability of medium-term momentum strategies in US equities (Jegadeesh 1990; Chan et al. 1996; and Jegadeesh and Titman 2001), the relation between price/earnings ratios and expected returns (Basu 1977), the volatility of orange juice futures prices (Roll 1984), and calendar effects such as holiday, weekend, and turn-of-the-month seasonalities (Lakonishok and Smidt 1988).

What are we to make of these anomalies? On the one hand, their persistence in the face of public scrutiny seems to be a clear violation of the EMH.

After all, most of these anomalies can be exploited by relatively simple trading strategies, and, while the resulting profits may not be riskless, they seem unusually profitable relative to their risks (see, especially, Lehmann 1990).

On the other hand, EMH supporters might argue that such persistence is in fact evidence in favour of EMH or, more to the point, that these anomalies cannot be exploited to any significant degree because of factors such as risk or transactions costs. Moreover, although some anomalies are currently inexplicable, this may be due to a lack of imagination on the part of academics, not necessarily a violation of the EMH. For example, recent evidence suggests that the January effect is largely due to 'bid–ask bounce', that is, closing prices for the last trading day of December tend to be at the bid price and closing prices for the first trading day of January tend to be at the ask price. Since small-capitalization stocks are also often low-price stocks, the effects of bid–ask bounce in percentage terms are much more pronounced for these stocks – a movement from bid to ask for a $5.00 stock on the NYSE (where the minimum bid-ask spread was $0.125 prior to decimalization in 2000) represents a 2.5 per cent return.

Whether or not one can profit from anomalies is a question unlikely to be settled in an academic setting. While calculations of 'paper' profits of various trading strategies come easily to academics, it is virtually impossible to incorporate in a realistic manner important features of the trading process such as transactions costs (including price impact), liquidity, rare events, institutional rigidities and non-stationarities. The economic value of anomalies must be decided in the laboratory of actual markets by investment professionals, over long periods of time, and even in these cases superior performance and simple luck are easily confused.

In fact, luck can play another role in the interpretation of anomalies: it can account for anomalies that are not anomalous. Regular patterns in historical data can be found even if no regularities exist, purely by chance. Although the likelihood of finding such spurious regularities is usually small (especially if the regularity is a very complex pattern), it increases dramatically with the number of 'searches' conducted on the same set of data. Such *data-snooping* biases are illustrated in Brown et al. (1992) and Lo and MacKinlay (1990b) – even the smallest biases can translate into substantial anomalies such as superior investment returns or the size effect.

Behavioural Critiques

The most enduring critiques of the EMH revolve around the preferences and behaviour of market participants. The standard approach to modelling preferences is to assert that investors optimize additive time-separable expected utility functions from certain parametric families – for example, constant relative risk aversion. However, psychologists and experimental economists have documented a number of departures from this paradigm, in the form of specific behavioural biases that are ubiquitous to human decision-making under uncertainty, several of which lead to undesirable outcomes for an individual's economic welfare – for example, overconfidence (Fischoff and Slovic 1980; Barber and Odean 2001; Gervais and Odean 2001), overreaction (DeBondt and Thaler 1985), loss aversion (Kahneman and Tversky 1979; Shefrin and Statman 1985; Odean 1998), herding (Huberman and Regev 2001), psychological accounting (Tversky and Kahneman 1981), miscalibration of probabilities (Lichtenstein et al. 1982), hyperbolic discounting (Laibson 1997), and regret (Bell 1982). These critics of the EMH argue that investors are often – if not always – irrational, exhibiting predictable and financially ruinous behaviour.

To see just how pervasive such behavioural biases can be, consider the following example which is a slightly modified version of an experiment conducted by two psychologists, Kahneman and Tversky (1979). Suppose you are offered two investment opportunities, A and B: A yields a sure profit of $240,000, and B is a lottery ticket yielding $1 million with a 25 per cent probability and $0 with 75 per cent probability. If you had to choose between A and *B,* which would you prefer? Investment *B* has an expected

value of $250,000, which is higher than A's payoff, but this may not be all that meaningful to you because you will receive either $1 million or zero. Clearly, there is no right or wrong choice here; it is simply a matter of personal preferences. Faced with this choice, most subjects prefer A, the sure profit, to B, despite the fact that B offers a significant probability of winning considerably more. This behaviour is often characterized as 'risk aversion' for obvious reasons. Now suppose you are faced with another two choices, C and D: C yields a sure loss of $750,000, and *D* is a lottery ticket yielding $0 with 25 per cent probability and a loss of $1 million with 75 per cent probability. Which would you prefer? This situation is not as absurd as it might seem at first glance; many financial decisions involve choosing between the lesser of two evils. In this case, most subjects choose D, despite the fact that D is more risky than C. When faced with two choices that both involve losses, individuals seem to be 'risk seeking', not risk averse as in the case of A versus B.

The fact that individuals tend to be risk averse in the face of gains and risk seeking in the face of losses can lead to some very poor financial decisions. To see why, observe that the combination of choices A and D is equivalent to a single lottery ticket yielding $240,000 with 25 per cent probability and − $760,000 with 75 per cent probability, whereas the combination of choices B and C is equivalent to a single lottery ticket yielding $250,000 with 25 per cent probability and − $750,000 with 75 per cent probability. The B and C combination has the same probabilities of gains and losses, but the gain is $10,000 higher and the loss is $10,000 lower. In other words, B and C is formally equivalent to A and D plus a sure profit of $10,000. In light of this analysis, would you still prefer A and D?

A common response to this example is that it is contrived because the two pairs of investment opportunities were presented sequentially, not simultaneously. However, in a typical global financial institution the London office may be faced with choices A and B and the Tokyo office may be faced with choices C and D. Locally, it may seem as if there is no right or wrong answer – the choice between A and B or C and D seems to be simply a matter of personal risk preferences – but the globally consolidated financial statement for the entire institution will tell a very different story. From that perspective, there *is* a right and wrong answer, and the empirical and experimental evidence suggests that most individuals tend to select the wrong answer. Therefore, according to the behaviouralists, quantitative models of efficient markets – all of which are predicated on rational choice – are likely to be wrong as well.

Impossibility of Efficient Markets

Grossman and Stiglitz (1980) go even farther – they argue that perfectly informationally efficient markets are an *impossibility* for, if markets are perfectly efficient, there is no profit to gathering information, in which case there would be little reason to trade and markets would eventually collapse. Alternatively, the degree of market *inefficiency* determines the effort investors are willing to expend to gather and trade on information, hence a non-degenerate market equilibrium will arise only when there are sufficient profit opportunities, that is, inefficiencies, to compensate investors for the costs of trading and information gathering. The profits earned by these attentive investors may be viewed as 'economic rents' that accrue to those willing to engage in such activities. Who are the providers of these rents? Black (1986) gave us a provocative answer: 'noise traders', individuals who trade on what they consider to be information but which is, in fact, merely noise.

The supporters of the EMH have responded to these challenges by arguing that, while behavioural biases and corresponding inefficiencies do exist from time to time, there is a limit to their prevalence and impact because of opposing forces dedicated to exploiting such opportunities. A simple example of such a limit is the so-called 'Dutch book', in which irrational probability beliefs give rise to guaranteed profits for the savvy investor. Consider, for example, an event *E,* defined as 'the S&P 500 index drops by five per cent or more next Monday', and suppose an individual has the following irrational beliefs: there is

a 50 per cent probability that E will occur, and a 75 per cent probability that E will *not* occur. This is clearly a violation of one of the basic axioms of probability theory – the probabilities of two mutually exclusive and exhaustive events must sum to 1 – but many experimental studies have documented such violations among an overwhelming majority of human subjects.

These inconsistent subjective probability beliefs imply that the individual would be willing to take both of the following bets B_1 and B_2:

$$B_1 = \begin{cases} \$1 \text{ if } E \\ -\$1 \text{ otherwise} \end{cases}, B_2 = \begin{cases} \$1 \text{ if } E^c \\ -\$1 \text{ otherwise} \end{cases}$$

where E^c denotes the event 'not E'. Now suppose we take the opposite side of both bets, placing \$50 on B_1 and \$25 on B_2. If E occurs, we lose \$50 on B_1 but gain \$75 on B_2, yielding a profit of \$25. If E^c occurs, we gain \$50 on B_1 and lose \$25 on B_2, also yielding a profit of \$25. Regardless of the outcome, we have secured a profit of \$25, an 'arbitrage' that comes at the expense of the individual with inconsistent probability beliefs. Such beliefs are not sustainable, and market forces – namely, arbitrageurs such as hedge funds and proprietary trading groups – will take advantage of these opportunities until they no longer exist, that is, until the odds are in line with the axioms of probability theory. (Only when these axioms are satisfied is arbitrage ruled out. This was conjectured by Ramsey 1926, and proved rigorously by de Finetti 1937, and Savage 1954.) Therefore, proponents of the classical EMH argue that there are limits to the degree and persistence of behavioural biases such as inconsistent probability beliefs, and substantial incentives for those who can identify and exploit such occurrences. While all of us are subject to certain behavioural biases from time to time, according to EMH supporters market forces will always act to bring prices back to rational levels, implying that the impact of irrational behaviour on financial markets is generally negligible and, therefore, irrelevant.

But this last conclusion relies on the assumption that market forces are sufficiently powerful to overcome any type of behavioural bias, or equivalently that irrational beliefs are not so pervasive as to overwhelm the capacity of arbitrage capital dedicated to taking advantage of such irrationalities. This is an empirical issue that cannot be settled theoretically, but must be tested through careful measurement and statistical analysis. The classic reference by Kindleberger (1989) – where a number of speculative bubbles, financial panics, manias, and market crashes are described in detail – suggests that the forces of irrationality can overwhelm the forces of arbitrage capital for months and, in several well-known cases, years.

So what does this imply for the EMH?

The Current State of the EMH

Given all of the theoretical and empirical evidence for and against the EMH, what can we conclude? Amazingly, there is still no consensus among economists. Despite the many advances in the statistical analysis, databases, and theoretical models surrounding the EMH, the main result of all of these studies is to harden the resolve of the proponents of each side of the debate.

One of the reasons for this state of affairs is the fact that the EMH, by itself, is not a well-defined and empirically refutable hypothesis. To make it operational, one must specify additional structure, for example, investors' preferences or information structure. But then a test of the EMH becomes a test of several auxiliary hypotheses as well, and a rejection of such a joint hypothesis tells us little about which aspect of the joint hypothesis is inconsistent with the data. Are stock prices too volatile because markets are inefficient, or due to risk aversion, or dividend smoothing? All three inferences are consistent with the data. Moreover, new statistical tests designed to distinguish among them will no doubt require auxiliary hypotheses of their own which, in turn, may be questioned.

More importantly, tests of the EMH may not be the most informative means of gauging the efficiency of a given market. What is often of more consequence is the efficiency of a particular market *relative* to other markets – for example, futures vs. spot markets, auction vs. dealer markets. The advantages of the concept of relative efficiency, as

opposed to the all-or-nothing notion of absolute efficiency, are easy to spot by way of an analogy. Physical systems are often given an efficiency rating based on the relative proportion of energy or fuel converted to useful work. Therefore, a piston engine may be rated at 60 per cent efficiency, meaning that on average 60 per cent of the energy contained in the engine's fuel is used to turn the crankshaft, with the remaining 40 per cent lost to other forms of work, such as heat, light or noise.

Few engineers would ever consider performing a statistical test to determine whether or not a given engine is perfectly efficient – such an engine exists only in the idealized frictionless world of the imagination. But measuring relative efficiency – relative, that is, to the frictionless ideal – is commonplace. Indeed, we have come to expect such measurements for many household products: air conditioners, hot water heaters, refrigerators, and so on. Therefore, from a practical point of view, and in light of Grossman and Stiglitz (1980), the EMH is an idealization that is economically unrealizable, but which serves as a useful benchmark for measuring relative efficiency.

The desire to build financial theories based on more realistic assumptions has led to several new strands of literature, including psychological approaches to risktaking behaviour (Kahneman and Tversky 1979; Thaler 1993; Lo 1999), evolutionary game theory (Friedman 1991), agent-based modelling of financial markets (Arthur et al. 1997; Chan et al. 1998), and direct applications of the principles of evolutionary psychology to economics and finance (Lo 1999, 2002, 2004, 2005; Lo and Repin 2002). Although substantially different in methods and style, these emerging sub-fields are all directed at new interpretations of the EMH. In particular, psychological models of financial markets focus on the manner in which human psychology influences the economic decision-making process as an explanation of apparent departures from rationality. Evolutionary game theory studies the evolution and steady-state equilibria of populations of competing strategies in highly idealized settings. Agent-based models are meant to capture complex learning behaviour and dynamics in financial markets using more realistic markets, strategies, and information structures. And applications of evolutionary psychology provide a reconciliation of rational expectations with the behavioural findings that often seem inconsistent with rationality.

For example, in one agent-based model of financial markets (Farmer 2002), the market is modelled using a non-equilibrium market mechanism, whose simplicity makes it possible to obtain analytic results while maintaining a plausible degree of realism. Market participants are treated as computational entities that employ strategies based on limited information. Through their (sometimes suboptimal) actions they make profits or losses. Profitable strategies accumulate capital with the passage of time, and unprofitable strategies lose money and may eventually disappear. A financial market can thus be viewed as a co-evolving ecology of trading strategies. The strategy is analogous to a biological species, and the total capital deployed by agents following a given strategy is analogous to the population of that species. The creation of new strategies may alter the profitability of pre-existing strategies, in some cases replacing them or driving them extinct.

Although agent-based models are still in their infancy, the simulations and related theory have already demonstrated an ability to understand many aspects of financial markets. Several studies indicate that, as the population of strategies evolves, the market tends to become more efficient, but this is far from the perfect efficiency of the classical EMH. Prices fluctuate in time with internal dynamics caused by the interaction of diverse trading strategies. Prices do not necessarily reflect 'true values'; if we view the market as a machine whose job is to set prices properly, the inefficiency of this machine can be substantial. Patterns in the price tend to disappear as agents evolve profitable strategies to exploit them, but this occurs only over an extended period of time, during which substantial profits may be accumulated and new patterns may appear.

The Adaptive Markets Hypothesis

The methodological differences between mainstream and behavioural economics suggest that an alternative to the traditional deductive approach of neoclassical economics may be necessary to reconcile the EMH with its behavioural critics. One particularly promising direction is to view financial markets from a biological perspective and, specifically, within an evolutionary framework in which markets, instruments, institutions and investors interact and evolve dynamically according to the 'law' of economic selection. Under this view, financial agents compete and adapt, but they do not necessarily do so in an optimal fashion (see Farmer and Lo 1999; Farmer 2002; Lo 2002, 2004, 2005).

This evolutionary approach is heavily influenced by recent advances in the emerging discipline of 'evolutionary psychology', which builds on the seminal research of E.O. Wilson (1975) in applying the principles of competition, reproduction, and natural selection to social interactions, yielding surprisingly compelling explanations for certain kinds of human behaviour, such as altruism, fairness, kin selection, language, mate selection, religion, morality, ethics and abstract thought (see, for example, Barkow et al. 1992; Gigerenzer 2000). 'Sociobiology' is the rubric that Wilson (1975) gave to these powerful ideas, which generated a considerable degree of controversy in their own right, and the same principles can be applied to economic and financial contexts. In doing so, we can fully reconcile the EMH with all of its behavioural alternatives, leading to a new synthesis: the adaptive markets hypothesis (AMH).

Students of the history of economic thought will no doubt recall that Thomas Malthus used biological arguments – the fact that populations increase at geometric rates whereas natural resources increase at only arithmetic rates – to arrive at rather dire economic consequences, and that both Darwin and Wallace were influenced by these arguments (see Hirshleifer 1977, for further details). Also, Joseph Schumpeter's view of business cycles, entrepreneurs and capitalism have an unmistakeable evolutionary flavour to them; in fact, his notions of 'creative destruction' and 'bursts' of entrepreneurial activity are similar in spirit to natural selection and Eldredge and Gould's (1972) notion of 'punctuated equilibrium'. More recently, economists and biologists have begun to explore these connections in several veins: direct extensions of sociobiology to economics (Becker 1976; Hirshleifer 1977); evolutionary game theory (Maynard Smith 1982); evolutionary economics (Nelson and Winter 1982); and economics as a complex system (Anderson et al. 1988). And publications like the *Journal of Evolutionary Economics* and the *Electronic Journal of Evolutionary Modeling and Economic Dynamics* now provide a home for research at the intersection of economics and biology.

Evolutionary concepts have also appeared in a number of financial contexts. For example, Luo (1995) explores the implications of natural selection for futures markets, and Hirshleifer and Luo (2001) consider the long-run prospects of overconfident traders in a competitive securities market. The literature on agent-based modelling pioneered by Arthur et al. (1997), in which interactions among software agents programmed with simple heuristics are simulated, relies heavily on evolutionary dynamics. And at least two prominent practitioners have proposed Darwinian alternatives to the EMH. In a chapter titled 'The Ecology of Markets', Niederhoffer (1997, ch. 15) likens financial markets to an ecosystem with dealers as 'herbivores', speculators as 'carnivores', and floor traders and distressed investors as 'decomposers'. And Bernstein (1998) makes a compelling case for active management by pointing out that the notion of equilibrium, which is central to the EMH, is rarely realized in practice and that market dynamics are better explained by evolutionary processes.

Clearly the time is now ripe for an evolutionary alternative to market efficiency. To that end, in the current context of the EMH we begin, as Samuelson (1947) did, with the theory of the individual consumer. Contrary to the neoclassical postulate that individuals maximize expected utility and have rational expectations, an evolutionary perspective makes considerably more modest claims, viewing individuals as organisms that

have been honed, through generations of natural selection, to maximize the survival of their genetic material (see, for example, Dawkins 1976). While such a reductionist approach can quickly degenerate into useless generalities – for example, the molecular biology of economic behaviour – nevertheless, there are valuable insights to be gained from the broader biological perspective. Specifically, this perspective implies that behaviour is not necessarily intrinsic and exogenous, but evolves by natural selection and depends on the particular environment through which selection occurs. That is, natural selection operates not only upon genetic material but also upon social and cultural norms in *Homo sapiens*; hence Wilson's term 'sociobiology'.

To operationalize this perspective within an economic context, consider the idea of 'bounded rationality' first espoused by Nobel-prize-winning economist Herbert Simon. Simon (1955) suggested that individuals are hardly capable of the kind of optimization that neoclassical economics calls for in the standard theory of consumer choice. Instead, he argued that, because optimization is costly and humans are naturally limited in their computational abilities, they engage in something he called 'satisficing', an alternative to optimization in which individuals make choices that are merely satisfactory, not necessarily optimal. In other words, individuals are bounded in their degree of rationality, which is in sharp contrast to the current orthodoxy – rational expectations – where individuals have unbounded rationality (the term 'hyperrational expectations' might be more descriptive). Unfortunately, although this idea garnered a Nobel Prize for Simon, it had relatively little impact on the economics profession. (However, his work is now receiving greater attention, thanks in part to the growing behavioural literature in economics and finance. See, for example, Simon 1982; Sargent 1993; Rubinstein 1998; Gigerenzer and Selten 2001.) Apart from the sociological factors discussed above, Simon's framework was commonly dismissed because of one specific criticism: what determines the point at which an individual stops optimizing and reaches a satisfactory solution? If such a point is determined by the usual cost–benefit calculation underlying much of microeconomics (that is, optimize until the marginal benefits of the optimum equals the marginal cost of getting there), this assumes the optimal solution is known, which would eliminate the need for satisficing. As a result, the idea of bounded rationality fell by the wayside, and rational expectations has become the de facto standard for modelling economic behaviour under uncertainty.

An evolutionary perspective provides the missing ingredient in Simon's framework. The proper response to the question of how individuals determine the point at which their optimizing behaviour is satisfactory is this: such points are determined not analytically but through trial and error and, of course, natural selection. Individuals make choices based on past experience and their 'best guess' as to what might be optimal, and they learn by receiving positive or negative reinforcement from the outcomes. If they receive no such reinforcement, they do not learn. In this fashion, individuals develop heuristics to solve various economic challenges, and, as long as those challenges remain stable, the heuristics will eventually adapt to yield approximately optimal solutions to them.

If, on the other hand, the environment changes, then it should come as no surprise that the heuristics of the old environment are not necessarily suited to the new. In such cases, we observe 'behavioural biases' – actions that are apparently ill–advised in the context in which we observe them. But rather than labelling such behaviour 'irrational', it should be recognized that suboptimal behaviour is not unlikely when we take heuristics out of their evolutionary context. A more accurate term for such behaviour might be 'maladaptive'. The flopping of a fish on dry land may seem strange and unproductive, but under water the same motions are capable of propelling the fish away from its predators.

By coupling Simon's notion of bounded rationality and satisficing with evolutionary dynamics, many other aspects of economic behaviour can also be derived. Competition, cooperation, market-making behaviour, general equilibrium, and disequilibrium dynamics are all adaptations designed to address certain environmental challenges for the human species, and by viewing them through the lens of evolutionary biology

we can better understand the apparent contradictions between the EMH and the presence and persistence of behavioural biases.

Specifically, the adaptive markets hypothesis can be viewed as a new version of the EMH, derived from evolutionary principles. Prices reflect as much information as dictated by the combination of environmental conditions and the number and nature of 'species' in the economy or, to use the appropriate biological term, the *ecology*. By 'species' I mean distinct groups of market participants, each behaving in a common manner. For example, pension funds may be considered one species; retail investors, another; market-makers, a third; and hedge-fund managers, a fourth. If multiple species (or the members of a single highly populous species) are competing for rather scarce resources within a single market, that market is likely to be highly efficient – for example, the market for 10-Year US Treasury Notes reflects most relevant information very quickly indeed. If, on the other hand, a small number of species are competing for rather abundant resources in a given market, that market will be less efficient – for example, the market for oil paintings from the Italian Renaissance. Market efficiency cannot be evaluated in a vacuum, but is highly context-dependent and dynamic, just as insect populations advance and decline as a function of the seasons, the number of predators and prey they face, and their abilities to adapt to an ever-changing environment.

The profit opportunities in any given market are akin to the amount of food and water in a particular local ecology – the more resources present, the less fierce the competition. As competition increases, either because of dwindling food supplies or an increase in the animal population, resources are depleted which, in turn, causes a population decline eventually, decreasing the level of competition and starting the cycle again. In some cases cycles converge to corner solutions, that is, certain species become extinct, food sources are permanently exhausted, or environmental conditions shift dramatically. By viewing economic profits as the ultimate food source on which market participants depend for their survival, the dynamics of market interactions and financial innovation can be readily derived.

Under the AMH, behavioural biases abound. The origins of such biases are heuristics that are adapted to non-financial contexts, and their impact is determined by the size of the population with such biases versus the size of competing populations with more effective heuristics. During the autumn of 1998, the desire for liquidity and safety by a certain population of investors overwhelmed the population of hedge funds attempting to arbitrage such preferences, causing those arbitrage relations to break down. However, in the years prior to August 1998 fixed-income relative-value traders profited handsomely from these activities, presumably at the expense of individuals with seemingly 'irrational' preferences (in fact, such preferences were shaped by a certain set of evolutionary forces, and might be quite rational in other contexts). Therefore, under the AMH, investment strategies undergo cycles of profitability and loss in response to changing business conditions, the number of competitors entering and exiting the industry, and the type and magnitude of profit opportunities available. As opportunities shift, so too will the affected populations. For example, after 1998 the number of fixed-income relative-value hedge funds declined dramatically – because of outright failures, investor redemptions, and fewer start-ups in this sector – but many have reappeared in recent years as performance for this type of investment strategy has improved.

Even fear and greed – the two most common culprits in the downfall of rational thinking according to most behaviouralists – are the product of evolutionary forces, adaptive traits that enhance the probability of survival. Recent research in the cognitive neurosciences and economics, now coalescing into the discipline known as 'neuroeconomics', suggests an important link between rationality in decision-making and emotion (Grossberg and Gutowski 1987; Damasio 1994; Elster 1998; Lo and Repin 2002; and Loewenstein 2000), implying that the two are not antithetical but in fact complementary. For example, contrary to the common belief that emotions have no place in rational financial decision-making processes, Lo and Repin (2002) present preliminary evidence that physiological variables associated with the autonomic nervous system are

highly correlated with market events even for highly experienced professional securities traders. They argue that emotional responses are a significant factor in the real-time processing of financial risks, and that an important component of a professional trader's skills lies in his or her ability to channel emotion, consciously or unconsciously, in specific ways during certain market conditions.

This argument often surprises economists because of the link between emotion and behavioural biases, but a more sophisticated view of the role of emotions in human cognition shows that they are central to rationality (see, for example, Damasio 1994; Rolls 1999). In particular, emotions are the basis for a reward-and-punishment system that facilitates the selection of advantageous behaviour, providing a numeraire for animals to engage in a 'cost–benefit analysis' of the various actions open to them (Rolls 1999, ch. 10.3). From an evolutionary perspective, emotion is a powerful adaptation that dramatically improves the efficiency with which animals learn from their environment and their past (see Damasio 1994). These evolutionary underpinnings are more than simple speculation in the context of financial market participants. The extraordinary degree of competitiveness of global financial markets and the outsize rewards that accrue to the 'fittest' traders suggest that Darwinian selection – 'survival of the richest', to be precise – is at work in determining the typical profile of the successful trader. After all, unsuccessful traders are eventually eliminated from the population after suffering a certain level of losses.

The new paradigm of the AMH is still under development, and certainly requires a great deal more research to render it 'operationally meaningful' in Samuelson's sense. However, even at this early stage it is clear that an evolutionary framework is able to reconcile many of the apparent contradictions between efficient markets and behavioural exceptions. The former may be viewed as the steady-state limit of a population with constant environmental conditions, and the latter involves specific adaptations of certain groups that may or may not persist, depending on the particular evolutionary paths that the economy experiences. More specific implications may be derived through a combination of deductive and inductive inference –for example, theoretical analysis of evolutionary dynamics, empirical analysis of evolutionary forces in financial markets, and experimental analysis of decision-making at the individual and group level.

For example, one implication is that, to the extent that a relation between risk and reward exists, it is unlikely to be stable over time. Such a relation is determined by the relative sizes and preferences of various populations in the market ecology, as well as institutional aspects such as the regulatory environment and tax laws. As these factors shift over time, any risk–reward relation is likely to be affected. A corollary of this implication is that the equity risk premium is also time-varying and path-dependent. This is not so revolutionary an idea as it might first appear – even in the context of a rational expectations equilibrium model, if risk preferences change over time, then the equity risk premium must vary too. The incremental insight of the AMH is that aggregate risk preferences are not immutable constants, but are shaped by the forces of natural selection. For example, until recently US markets were populated by a significant group of investors who had never experienced a genuine bear market – this fact has undoubtedly shaped the aggregate risk preferences of the US economy, just as the experience since the bursting of the technology bubble in the early 2000s has affected the risk preferences of the current population of investors. In this context, natural selection determines who participates in market interactions; those investors who experienced substantial losses in the technology bubble are more likely to have exited the market, leaving a markedly different population of investors. Through the forces of natural selection, history matters. Irrespective of whether prices fully reflect all available information, the particular path that market prices have taken over the past few years influences current aggregate risk preferences. Among the three fundamental components of any market equilibrium – prices, probabilities, and preferences – preferences is clearly the most fundamental and least understood. Several large bodies of research have developed around these issues – in economics

and finance, psychology, operations research (also called 'decision sciences') and, more recently, brain and cognitive sciences – and many new insights are likely to flow from synthesizing these different strands of research into a more complete understanding of how individuals make decisions (see Starmer 2000, for an excellent review of this literature). Simon's (1982) seminal contributions to this literature are still remarkably timely and their implications have yet to be fully explored.

Conclusions

Many other practical insights and potential breakthroughs can be derived from shifting our mode of thinking in financial economics from the physical to the biological sciences. Although evolutionary ideas are not yet part of the financial mainstream, the hope is that they will become more commonplace as they demonstrate their worth – ideas are also subject to 'survival of the fittest'. No one has illustrated this principal so well as Harry Markowitz, the father of modern portfolio theory and a Nobel laureate in economics in 1990. In describing his experience as a Ph.D. student on the eve of his graduation, he wrote in his Nobel address (Markowitz 1991, p. 476):

> . . . [W]hen I defended my dissertation as a student in the Economics Department of the University of Chicago, Professor Milton Friedman argued that portfolio theory was not Economics, and that they could not award me a Ph.D. degree in Economics for a dissertation which was not Economics. I assume that he was only half serious, since they did award me the degree without long debate. As to the merits of his arguments, at this point I am quite willing to concede: at the time I defended my dissertation, portfolio theory was not part of Economics. But now it is.

In light of the sociology of the EMH controversy (see, for example, Lo 2004), the debate is likely to continue. However, despite the lack of consensus in academia and industry, the ongoing dialogue has given us many new insights into the economic structure of financial markets. If, as Paul Samuelson has suggested, financial economics is the crown jewel of the social sciences, then the EMH must account for half the facets.

See Also

- ▶ Financial Market Anomalies
- ▶ Rational Expectations
- ▶ Rationality, Bounded

Acknowledgment I thank John Cox, Gene Fama, Bob Merton, and Paul Samuelson for helpful discussions.

Bibliography

Anderson, P., K. Arrow, and D. Pines, eds. 1988. *The economy as an evolving complex system*. Reading: Addison-Wesley Publishing Company.

Arthur, B., J. Holland, B. LeBaron, R. Palmer, and P. Tayler. 1997. Asset pricing under endogenous expectations in an artificial stock market. In *The economy as an evolving complex system II*, ed. B. Arthur, S. Durlauf, and D. Lane. Reading: Addison Wesley.

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6: 159–178.

Banz, R. 1981. The relationship between return and market value of common stock. *Journal of Financial Economics* 9: 3–18.

Barber, B., and T. Odean. 2001. Boys will be boys: gender, overconfidence, and common stock investment. *Quarterly Journal of Economics* 116: 261–229.

Barkow, J., L. Cosmides, and J. Tooby. 1992. *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford: Oxford University Press.

Basu, S. 1977. The investment performance of common stocks in relation to their price–earnings ratios: A test of the efficient market hypothesis. *Journal of Finance* 32: 663–682.

Becker, G. 1976. Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature* 14: 817–826.

Bell, D. 1982. Risk premiums for decision regret. *Management Science* 29: 1156–1166.

Bernard, V., and J. Thomas. 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics* 13: 305–340.

Bernstein, P. 1998. Why the efficient market offers hope to active management. In *Economics and portfolio strategy*. New York: Peter Bernstein, Inc, October 1.

Black, F. 1986. Noise. *Journal of Finance* 41: 529–544.

Black, F., and M. Scholes. 1973. Pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Brown, S., W. Goetzmann, R. Ibbotson, and S. Ross. 1992. Survivorship bias in performance studies. *Review of Financial Studies* 5: 553–580.

Campbell, J., and R. Shiller. 1988. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1: 195–228.

Chan, K. 1988. On the contrarian investment strategy. *Journal of Business* 61: 147–164.

Chan, L., N. Jegadeesh, and J. Lakonishok. 1996. Momentum strategies. *Journal of Finance* 51: 1681–1713.

Chan, N., B. LeBaron, A. Lo, and T. Poggio. 1998. *Information dissemination and aggregation in asset markets with simple intelligent traders*, Laboratory Technical Memorandum No. 1646. Cambridge, MA: MIT Artificial Intelligence.

Chopra, N., J. Lakonishok, and J. Ritter. 1992. Measuring abnormal performance: Do stocks overreact? *Journal of Financial Economics* 31: 235–286.

Cootner, P. 1962. Stock prices: Random vs. systematic changes. *Industrial Management Review* 3: 24–45.

Cootner, P. 1964. *The random character of stock market prices*. London: Risk Publications.

Copeland, T., and D. Mayers. 1982. The value line enigma (1965–1978): A case study of performance evaluation issues. *Journal of Financial Economics* 10: 289–322.

Cowles, A. 1960. A revision of previous conclusions regarding stock price behavior. *Econometrica* 28: 909–915.

Cowles, A., and H. Jones. 1937. Some a posteriori probabilities in stock market action. *Econometrica* 5: 280–294.

Damasio, A. 1994. *Descartes' error: Emotion, reason, and the human brain*. New York: Avon Books.

Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford University Press.

de Finetti, B. 1937. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré* 7: 1–68. English translation In *Studies in subjective probability*, ed. H. Kyburg and H. Smokler. New York: Wiley, 1964.

DeBondt, W., and R. Thaler. 1985. Does the stock market overreact? *Journal of Finance* 40: 793–807.

Eldredge, N., and S. Gould. 1972. Punctuated equilibria: An alternative to phyletic gradualism. In *Models in paleobiology*, ed. T. Schopf. San Francisco: Freeman/Cooper.

Elster, J. 1998. Emotions and economic theory. *Journal of Economic Literature* 36: 47–74.

Fama, E. 1963. Mandelbrot and the stable paretian hypothesis. *Journal of Business* 36: 420–429.

Fama, E. 1965a. The behavior of stock market prices. *Journal of Business* 38: 34–105.

Fama, E. 1965b. Random walks in stock market prices. *Financial Analysts Journal* 21: 55–59.

Fama, E. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.

Fama, E., and M. Blume. 1966. Filter rules and stock market trading profits. *Journal of Business* 39: 226–241.

Fama, E., and K. French. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* 96: 246–273.

Farmer, D. 2002. Market force, ecology and evolution. *Industrial and Corporate Change* 11: 895–953.

Farmer, D., and A. Lo. 1999. Frontiers of finance: Evolution and efficient markets. *Proceedings of the National Academy of Sciences* 96: 9991–9992.

Fischoff, B., and P. Slovic. 1980. A little learning. . .: Confidence in multicue judgment tasks. In *Attention and performance, VIII*, ed. R. Nickerson. Hillsdale: Erlbaum.

Flavin, M. 1983. Excess volatility in the financial markets: A reassessment of the empirical evidence. *Journal of Political Economy* 91: 929–956.

French, K., and R. Roll. 1986. Stock return variances: The arrival of information and the reaction of traders. *Journal of Financial Economics* 17: 5–26.

Friedman, D. 1991. Evolutionary games in economics. *Econometrica* 59: 637–666.

Gervais, S., and T. Odean. 2001. Learning to be overconfident. *Review of Financial Studies* 14: 1–27.

Gigerenzer, G. 2000. *Adaptive thinking: rationality in the real world*. Oxford: Oxford University Press.

Gigerenzer, G., and R. Selten. 2001. *Bounded rational: The adaptive toolbox*. Cambridge, MA: MIT Press.

Gilles, C., and S. LeRoy. 1991. Econometric aspects of the variance-bounds tests: A survey. *Review of Financial Studies* 4: 753–792.

Grossberg, S., and W. Gutowski. 1987. Neural dynamics of decision making under risk: Affective balance and cognitive-emotional interactions. *Psychological Review* 94: 300–318.

Grossman, S., and R. Shiller. 1981. The determinants of the variability of stock market prices. *American Economic Review* 71: 222–227.

Grossman, S., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Hirshleifer, J. 1977. Economics from a biological viewpoint. *Journal of Law and Economics* 20: 1–52.

Hirshleifer, D., and G. Luo. 2001. On the survival of overconfident traders in a competitive securities market. *Journal of Financial Markets* 4: 73–84.

Huberman, G., and T. Regev. 2001. Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *Journal of Finance* 56: 387–396.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *Journal of Finance* 45: 881–898.

Jegadeesh, N., and S. Titman. 2001. Profitability of momentum strategies: An evaluation of alternative explanations. *Journal of Finance* 56: 699–720.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Keim, D. 1983. Size-related anomalies and stock return seasonality: Further empirical evidence. *Journal of Financial Economics* 12: 13–32.

Kim, M., C. Nelson, and R. Startz. 1991. Mean reversion in stock prices? A reappraisal of the empirical evidence. *Review of Economic Studies* 58: 515–528.

Kindleberger, C. 1989. *Manias, panics, and crashes: A history of financial crises*. New York: Basic Books.

Kleidon, A. 1986. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94: 953–1001.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 62: 443–477.

Lakonishok, J., and S. Smidt. 1988. Are seasonal anomalies real? A ninety-year perspective. *Review of Financial Studies* 1: 403–425.

Lehmann, B. 1990. Fads, martingales, and market efficiency. *Quarterly Journal of Economics* 105: 1–28.

Leroy, S. 1973. Risk aversion and the martingale property of stock returns. *International Economic Review* 14: 436–446.

LeRoy, S., and R. Porter. 1981. The present value relation: Tests based on variance bounds. *Econometrica* 49: 555–574.

Lichtenstein, S., B. Fischoff, and L. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. In *Judgment under uncertainty: Heuristics and biases*, ed. D. Kahneman, P. Slovic, and A. Tversky. Cambridge: Cambridge University Press.

Lo, A. 1991. Long-term memory in stock market prices. *Econometrica* 59: 1279–1313.

Lo, A. (ed.). 1997. *Market efficiency: Stock market behavior in theory and practice,* 2 vols. Cheltenham: Edward Elgar Publishing Company.

Lo, A. 1999. The three P's of total risk management. *Financial Analysts Journal* 55: 87–129.

Lo, A. 2001. Risk management for hedge funds: Introduction and overview. *Financial Analysts Journal* 57: 16–33.

Lo, A. 2002. Bubble, rubble, finance in trouble? *Journal of Psychology and Financial Markets* 3: 76–86.

Lo, A. 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management* 30: 15–29.

Lo, A. 2005. Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis. *Journal of Investment Consulting* 7: 21–44.

Lo, A., and C. MacKinlay. 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* 1: 41–66.

Lo, A., and C. MacKinlay. 1990a. An econometric analysis of nonsynchronous trading. *Journal of Econometrics* 45: 181–212.

Lo, A., and C. MacKinlay. 1990b. Data snooping biases in tests of financial asset pricing models. *Review of Financial Studies* 3: 431–468.

Lo, A., and C. MacKinlay. 1990c. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3: 175–206.

Lo, A., and C. MacKinlay. 1999. *A non-random walk down wall street*. Princeton: Princeton University Press.

Lo, A., and D. Repin. 2002. The psychophysiology of real-time financial risk processing. *Journal of Cognitive Neuroscience* 14: 323–339.

Loewenstein, G. 2000. Emotions in economic theory and economic behavior. *American Economic Review* 90: 426–432.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1446.

Luo, G. 1995. Evolution and market competition. *Journal of Economic Theory* 67: 223–250.

Markowitz, H. 1991. Foundations of portfolio theory. *Journal of Finance* 46: 469–477.

Marsh, T., and R. Merton. 1986. Dividend variability and variance bounds tests for the rationality of stock market prices. *American Economic Review* 76: 483–498.

Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.

Merton, R. 1987. On the current state of the stock market rationality hypothesis. In *Macroeconomics and finance: Essays in honor of Franco Modigliani*, ed. R. Dornbusch, S. Fischer, and J. Bossons. Cambridge, MA: MIT Press.

Michener, R. 1982. Variance bounds in a simple model of asset pricing. *Journal of Political Economy* 90: 166–175.

Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Belknap Press of Harvard University Press.

Niederhoffer, V. 1997. *Education of a speculator*. New York: Wiley.

Odean, T. 1998. Are investors reluctant to realize their losses? *Journal of Finance* 53: 1775–1798.

Osborne, M. 1959. Brownian motion in the stock market. *Operations Research* 7: 145–173.

Poterba, J., and L. Summers. 1988. Mean reversion in stock returns: Evidence and implications. *Journal of Financial Economics* 22: 27–60.

Ramsey, F. 1926. Truth and probability. In *Foundations of mathematics and other logical essays*, ed. R. Braithwaite. New York: Harcourt Brace & Co.

Richardson, M. 1993. Temporary components of stock prices: A skeptic's view. *Journal of Business and Economics Statistics* 11: 199–207.

Roberts, H. 1959. Stock-market 'patterns' and financial analysis: Methodological suggestions. *Journal of Finance* 14: 1–10.

Roberts, H. 1967. *Statistical versus clinical prediction of the stock market*. Unpublished manuscript, Center for Research in Security Prices, University of Chicago, Chicago.

Roll, R. 1983. Vas is das? The turn-of-the-year effect and the return premia of small firms. *Journal of Portfolio Management* 9: 18–28.

Roll, R. 1984. Orange juice and weather. *American Economic Review* 74: 861–880.

Rolls, E. 1999. *The brain and emotion*. Oxford: Oxford University Press.

Rosenberg, B., K. Reid, and R. Lanstein. 1985. Persuasive evidence of market inefficiency. *Journal of Portfolio Management* 11: 9–17.

Rozeff, M., and W. Kinney Jr. 1976. Capital market seasonality: The case of stock returns. *Journal of Financial Economics* 3: 379–402.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7: 407–425.

Rubinstein, A. 1998. *Modeling bounded rationality*. Cambridge, MA: MIT Press.
Samuelson, P. 1947. *Foundations of economics analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.
Sargent, T. 1993. *Bounded rationality in macroeconomics*. Oxford: Clarendon Press.
Savage, L. 1954. *Foundations of statistics*. New York: Wiley.
Schumpeter, J. 1939. *Business cycles: A theoretical, historical, and statistical analysis of the capitalist process*. New York: McGraw-Hill.
Shefrin, M., and M. Statman. 1985. The disposition to sell winners too early and ride losers too long: theory and evidence. *Journal of Finance* 40: 777–790.
Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.
Simon, H. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69: 99–118.
Simon, H. 1982. *Models of bounded rationality*, 2 vols. Cambridge, MA: MIT Press.
Starmer, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38: 332–382.
Thaler, R., ed. 1993. *Advances in behavioral finance*. New York: Russell Sage Foundation.
Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211: 453–458.
Wilson, E. 1975. *Sociobiology: The new synthesis*. Cambridge, MA: Belknap Press of Harvard University Press.

Egalitarianism

Harry Brighouse and Adam Swift

Abstract

This article surveys a variety of egalitarian theories. We look at a series of different answers to the question of what the metric of justice should be. Then we survey different interpretations of the egalitarian distributive rule, including 'equality', 'prioritizing benefit to the least advantaged' and 'sufficiency'. Theories also differ by whether they see equality as properly holding within social institutions or being a principle that applies more cosmically. Finally, we observe that egalitarian theories differ as to the weight they grant to egalitarian values relative to other values.

Keywords

Compensation; Distributive justice; Dworkin, R; Egalitarianism; Envy; Equality; Equality of capabilities; Equality of opportunity; Equality of resources; Happiness; Interpersonal well-being comparisons; Justice; Luck egalitarianism; Opportunity cost; Positional goods; Prioritarianism; Property rights; Rawls, J; Relational egalitarianism; Sen, A; Sufficientarianism; Value pluralism; Welfare; Well-being

JEL Classification

D6

All modern political theories assume that persons are in some relevant sense moral equals, entitled to equal concern, respect or treatment, and that a theory of justice must interpret and reflect that moral equality. This commitment is sometimes dubbed the 'egalitarian plateau', and it has been a common foundational moral assumption since Locke. Contemporary theories differ in how they *interpret* the egalitarian plateau. Two kinds of theory of justice are usually counted as egalitarian. Theories of distributive equality concern themselves with the relative standing of individuals in the distribution of benefits and burdens; theories of relational equality concern themselves with the relative standing of individuals when they face each other in the public sphere.

The Metric

One key question concerns the metric of equality: what, precisely, is it that egalitarians should seek to equalize? The literature falls into three main camps. Resourcists argue that people should be equal in the space of resources, meaning that they should have equal opportunity for achieving holdings of alienable goods. How are holdings priced? Ronald Dworkin imagines a hypothetical auction in which persons with equal holdings of some

currency bid for available goods until markets clear (Dworkin 2000). The distribution after the auction is equal if no one prefers anyone else's bundle of goods to her own; the distribution is then said to pass the 'envy test'. The intuitive idea is that the price of some good is set by the opportunity cost to others of that good. We have to tailor our preferences to our resources; equality is achieved when all face the same budget constraint, not when all achieve equal satisfaction.

Equality of resources has difficulty with the intuition that those with less socially valued talent, and in particular those with serious impairments, should receive compensation. Two strategies are available. One is to adopt a view that talent is socially constructed, so that much of the disadvantage faced by the less talented and the impaired is a consequence not of their lack of talent but of the fact that social institutions are maladapted to their natural endowments (Pogge 2003). This view allows resourcists to call for the reform of social institutions in the name of equality, without demanding compensation for impairments. The problem with this strategy is that some mental and physical impairments *intrinsically* cause disadvantage; there is no feasible set of social arrangements that would not make it more difficult for people with the impairments to derive satisfaction from resources. So an alternative strategy is to make the cut between persons and resources in a different place, regarding talents as resources and disabilities as resource-deficits. Dworkin's own version of this strategy proposes compensating the less talented with additional income, the amount calculated by looking at the insurance that talented individuals would have bought against a lack of talents if they had no knowledge of their probability of having the talents.

An alternative metric is welfare; egalitarians of welfare would seek to equalize levels of welfare (understood sometimes as idealized preference satisfaction, sometimes in terms of internal states such as happiness). This view handles talent-inequality in a straightforward manner; the less talented and the disabled should be compensated up to the level where they enjoy as much welfare as anyone else. But it faces the problem that there is no reason for people to moderate their preferences; since welfare is a direct target, those with expensive tastes receive more resources than those with inexpensive tastes, which is widely regarded as intuitively unfair. An alternative view – equality of opportunity for welfare – deals with this problem by seeking equality of welfare except when inequalities are the result of voluntary well-informed choices rather than bad luck or circumstances outside the agent's control (Arneson 1989). Again, the less talented are straightforwardly compensated for the way in which they find it harder than others to derive satisfaction, but those who cultivate expensive tastes are not. However, those with non-cultivated expensive preferences are also compensated, even if they could easily be overcome; this view does not see lack of talent, and disability in particular, as morally more urgent than expensive preferences. (See Roemer 1986, for an argument that equality of resources implies equality of welfare.)

All of the views deploying an 'opportunity' metric, including Dworkin's resourcist view, presume the desirability of holding people accountable for their voluntary choices, but compensating them for deficits that are beyond their control. Views of this kind are sometimes referred to as varieties of 'luck egalitarianism'. Inequalities resulting from voluntary choice are acceptable because they reflect a deeper sense in which we are equal as moral agents; choice legitimizes inequality, brute luck does not. (For an elegant attempt both to conceptualize and operationalize equality of opportunity *tout court*, see Roemer 1998.)

The main rival account – namely, the capabilities approach developed by Amartya Sen and Martha Nussbaum – focuses on the *preconditions* of agency (Sen 1999; Nussbaum 2000). Equality of capabilities demands that people be equal in the space of the functionings or livings that they are substantively able to achieve. Walking is a functioning, so are eating, reading, mountain climbing, and chatting. 'The concept of functionings… reflects the various things a person may value doing or being – varying from the basic (being adequately nourished) to the very

complex (being able to take part in the life of the community)' (Sen 1999, p. 75). But when we make interpersonal comparisons of well-being we should find a measure that incorporates references to functionings but also reflects the intuition that what matters is not merely achieving the functioning but being free to achieve it. So we should look at 'the freedom to achieve actual livings that one can have a reason to value' (Sen 1999, p. 73) or, to put it another way, 'substantive freedoms – the capabilities – to choose a life one has reason to value'. The idea is that people should be equal in this space.

The capabilities approach avoids the problems of the standard welfarist approaches by focusing on *choice* (thus treating inequalities arising from voluntary choices differently from those arising from circumstances). It avoids the difficulty resourcist accounts have with unequal talent by focusing on *functionings*; talent deficits are compensated for by looking not at what others would pay to avoid them but at the valuable activities the deficits deprive people of access to. Some theorists place the capabilities account in the welfarist camp (Williams 2002) but it is not implausible to think of it as a variant of resourcism, distinguished by its approach to the valuation of talents.

A major recent development in the debates about egalitarianism has involved criticisms of luck egalitarianism. Each of the luck egalitarian principles, taken alone, imposes heavy costs on those who endure misfortunes for which they can be held responsible, even if those costs place the agent below the threshold for full participation in social affairs. An alternative has developed which is best described as 'relational egalitarianism'. Relational egalitarianism is not directly concerned with equality in terms of the distribution of any particular currency, but endorses the idea that individuals should have equal standing in the public sphere. This vague idea has several instantiations. Elizabeth Anderson (1999, p. 304) talks of seeking 'a social order in which persons stand in relations of equality'; Nancy Fraser (1998, p. 30) says that 'Justice requires social arrangements that permit all (adult) members of society to interact with one another as peers'. Both fill out their theories with more details. According to Fraser (1998, p. 24), 'It is unjust that some individuals and groups are denied the status of full partners in social interaction, simply as a consequence of institutionalized patterns of interpretation and evaluation in whose construction they have not equally participated and that disparage their distinctive characteristics or the distinctive characteristics assigned to them'. A third variant of relational egalitarianism spells it out specifically in terms of political equality, the idea being that it is particularly important that people enjoy equal availability of or opportunity for political power or influence (Christiano 1995). This variant is typically less hostile than other variants to luck egalitarianism.

Each of the views reviewed in this section allows inequality along some dimensions. Relational egalitarianisms allow such inequalities of income, wealth, welfare or capabilities as are compatible with equal political influence, or interaction as peers, or 'equal opportunity for participation as a peer'. These permitted inequalities may be great or very small, and how great or small may vary by social context. Principles demanding equality of opportunity are consistent with great inequalities in outcome, and consistent also with some being very badly off in absolute terms. While equality of opportunity conceptions place no limit on how badly off someone may be as a result of her own imprudent choices, equality of social standing demands that no one fall below the threshold needed for equal participation, even if she makes numerous imprudent choices.

The Distributive Rules

Do egalitarians even care about *equality*? Principles demanding equality of X seem vulnerable to an obvious objection. In some dynamic situations it is possible to produce more of X by distributing X unequally, and to ensure that even those with least have more than under an equal distribution. For example, we can sometimes produce more wealth by judiciously attaching higher income to more productive positions in the economy, and to

longer work hours; the higher income acts both as a signal and as an incentive to produce more. That greater production can be turned to the benefit of those with least. But, the objection goes, it would be perverse to prefer an equal situation in which everyone has less to one in which everyone has more, even if we have to sacrifice equality for the sake of that additional product.

This is known as the 'levelling down' objection to equality. Egalitarians make two distinct responses. The first is to concede the argument, abandoning 'equality' and replacing it with 'giving priority to the interests of the least advantaged'. John Rawls's difference principle, which states that 'social and economic inequalities are to be arranged to the maximum benefit of the least advantaged', embodies one variant of this response, a variant that gives *absolute* priority to the prospects of the least advantaged (Rawls 1971, 2001). A weaker variant in this family of views, usually known as 'prioritarianism', simply says that it is more urgent to provide benefits to those with less advantage than to those with more (Parfit 2000).

An alternative response is to assert value pluralism. This response acknowledges that priority to the least advantaged is an important value and perhaps more important than equality, so that when it comes to policy or action prioritarian principles should govern. But it says that equality nevertheless matters some; there is one way in which an unequal distribution is worse than an equal distribution, even if, all things considered, it is better; the way in which it is worse is that it is unequal and for that reason unfair (Temkin 2002). This response is bolstered by the observation that there is nothing eccentric about endorsing a principle that values distributions that benefit nobody; the retributive principle of proportionality between punishment and crime, for example, calls for harming the criminal even when there is no gain to anyone else in harming him.

Some reject principles of equality and priority on the grounds that all that matters for the purposes of justice is that all have enough. Sufficientarian theories are not usually counted as within the egalitarian family, because they eschew any fundamental concern with relativities. Relativities may matter in determining what is enough for people to live a decent life in any given social environment, but ultimately what matters is not where someone ranks in the distribution of resources (or anything else) but whether she has enough. However, as suggested above, sufficientarian principles also have a place in some variants of egalitarianism. While relational egalitarianism places no principled limits on the level of material or welfare inequality, and gives no general priority to the least advantaged, it does set a floor – all must have sufficient resources to be full participants in social interaction. Equality of political influence demands that all have sufficient resources, personal and financial, to play an equal role in political life, but, as long as it is possible to insulate politics from residual inequalities of wealth, it is not concerned with equalizing or prioritizing benefit to the least advantaged.

Many theories of justice that do not fit the above characterizations of egalitarianism nevertheless incorporate some elements of egalitarian thinking. John Rawls's theory of justice, for example, prioritizes the principle that certain basic liberties (not including strong property rights) be equally distributed, then demands that within that constraint fair equality of opportunity should be implemented, and then that social and economic inequalities be arranged to the greatest benefit of the least advantaged in so far as that is possible without jeopardizing the equal liberty and fair equality of opportunity principle (Rawls 1971, 2001). Michael Walzer's (1983) theory of 'complex equality' takes seriously widely shared intuitions that different goods are subject to different distributive rules. For example, while income should be distributed according to productive contribution, as will tend to result from market interactions, the inequalities this norm generates should be prevented from translating into unequal access to certain key goods like health care and educational opportunities, the distribution of which should be governed by need and the requirements of equal opportunity respectively. It is unclear in what sense Walzer's 'complex equality' is genuinely an egalitarian position, since it is in principle consistent with unequal

and coinciding distributions of all goods that are not themselves governed by egalitarian norms.

Priority and equality coincide in practice for one class of goods: positional goods. These have the property that the contribution an individual's share of the good makes to her absolute position is determined by how much of the good she has relative to others. The credentialing aspect of education is a paradigm case; how useful a degree is in landing a job (as opposed to the learning one achieved in the process of getting the degree) depends entirely on the credentials of one's competitors for that job. (Other cases are detailed in Hirsch 1976.) Those who give priority to the worst-off will countenance inequalities in positional goods only in so far as they are required by or result in the least advantaged benefiting overall (Brighouse and Swift 2006).

The Scope of Equality

Whatever the right distribuendum, and whatever the appropriate distributive principle, it is a further question who should be equal to whom. Some limit the application of their egalitarianism to members of the same society or system of cooperation, or to those subject to the same coercive structure (Nagel 2005), or hold that it is states that owe their citizens a particular duty to treat them with equal concern and respect (Dworkin 2000). Others believe that egalitarian principles should apply to all human beings, irrespective of the relations that obtain between them. If we restrict the application of egalitarian principles to schemes of cooperation, that does not exclude the possibility of a global egalitarianism, since most now accept that in the modern world social cooperation extends well beyond national boundaries (Julius 2006). But consider this version of Derek Parfit's divided world case. All the people in A are half as well off as all the people in B, but A and B have no knowledge of or contact with each other (Parfit 2000). Is there anything regrettable from the perspective of injustice about this inequality? If so, then the scope of justice is cosmic, not simply social. In the stated version of the divided world case this difference is motivationally inert, since the people in B do not have the relevant knowledge. But, if they did, cosmic egalitarianism would give them a reason to try to find a way to contact and interact with the people on A, while intra-societal egalitarianism would provide them with no such reason.

The divided world case brings out another difference in orientation. Where members of A and B have no interaction, or even knowledge of each other, equality can be valued only intrinsically rather than because of its effects on members of A or Often, however, inequality with respect to some goods is devalued, and equality valued, instrumentally, because of its absolute effects on those subject to the unequal distribution – usually its effects on the relatively disadvantaged. Thus, for example, economic inequalities are thought to undermine the fairness of legal or political processes, or occupational or other status hierarchies are claimed to harm the health of those on the lower rungs. Those who value equality intrinsically would hold that there is a reason to level down for the sake of equality or fairness, whereas instrumental egalitarians might seek the more equal distribution of some goods, not for egalitarian reasons *stricto sensu*, but to eliminate the bad effects of certain kinds of inequality.

The Subject of Justice

A further dividing line between egalitarians concerns the subject of justice. Rawls stipulates that the subject is the 'basic structure of society', which consists of some of the central, interaction-shaping institutions of a society: for example, the constitution, the legally recognized forms of property, the structure of the economy, the design of the legislature, and the judiciary. The idea is that these institutions govern the division of the advantages that accrue from social cooperation, and they assign the basic rights and responsibilities to citizens. So a society is just when those institutions are arranged according to the correct principles.

Rawls officially exempts individual actions and motives from evaluation from the perspective of egalitarian justice, as long as individuals obey the rules set by a just basic structure. But this has

the consequence that a society in which talented individuals take advantage of the prerogatives not to serve the least advantaged that are built into the principles that he thinks justice requires of coercive institutions is no less just than one in which they are much more strongly motivated by the desire to benefit the least advantaged through their choices regarding work. A society with an egalitarian governing ethos, on this view, is no more just than one without, even when the least advantaged are much better off. But the motivations and actions of talented individuals affect the prospects and status of others in ways that have 'profound and pervasive influence on persons' (Rawls 2001, p. 55), which is Rawls's central reason for focusing on the basic structure. So some egalitarians regard justice as commenting not only on the broad coercive outline of society, but also on less officially coercive institutions such as a society's ethos (Cohen 1997). For a powerful defence of an account intermediate between Cohen's and Rawls's, see Julius 2003).

Other Values

Most egalitarian theorists are value pluralists; they believe that equality (or priority) of their preferred metric matters, but so do other principles. Observing that equality or priority is sometimes in conflict with liberty or privacy or efficiency does not require us to reject one of the conflicting values. It requires us, instead, to evaluate reasons for considering one of the values more morally important than the others, and, in the light of that evaluation, to establish which should give way in different conflicts. Unless the relationship between values is one of lexical priority (in which case the prior value always trumps subordinate values, which can be pursued only when there is no conflict), different trade-offs between values will be mandated in different conflicts. But lexical priority is unlikely to hold between genuine values. If a value matters *at all*, it is hard to believe it could never be the case that a very large amount of it was greater than a very small amount of a conflicting value *however great that conflicting value is*.

See Also

- ▶ Equality of Opportunity
- ▶ Ethics and Economics
- ▶ Liberalism and Economics
- ▶ Libertarianism
- ▶ Pareto Efficiency
- ▶ Satisficing
- ▶ Sen, Amartya (Born 1933)

Bibliography

Anderson, E. 1999. What is the point of equality? *Ethics* 109: 287–337.

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56: 77–93.

Arneson, R. 2002. Egalitarianism. In: *Stanford encyclopedia of philosophy*. Online. Available at http://plato.stanford.edu/entries/egalitarianism. Accessed 14 Oct 2006.

Brighouse, H., and A. Swift. 2006. Equality, priority and positional goods. *Ethics* 116: 471–497.

Christiano, T. 1995. *The rule of the many*. Boulder: Westview Press.

Cohen, G.A. 1997. Where the action is: On the site of distributive justice. *Philosophy and Public Affairs* 26: 3–30.

Dworkin, R. 1981. What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs* 10: 283–345.

Dworkin, R. 2000. *Sovereign virtue*. Cambridge, MA: Harvard University Press.

Fraser, N. 1998. Social justice in the age of identity politics. In: *The Tanner lectures on human values*. Stanford University, 29 April–2 May. Online. Available at http://www.tannerlectures.utah.edu/lectures/Fraser98.pdf. Accessed 14 Oct 2006.

Hirsch, F. 1976. *Social limits to growth*. London: Routledge & Kegan Paul.

Julius, A.J. 2003. Basic structure and the value of equality. *Philosophy and Public Affairs* 31: 321–355.

Julius, A.J. 2006. Nagel's atlas. *Philosophy and Public Affairs* 34: 176–192.

Nagel, T. 2005. The problem of global justice. *Philosophy and Public Affairs* 33: 113–147.

Nussbaum, M. 2000. *Women and human development*. Cambridge: Cambridge University Press.

Parfit, D. 2000. Equality or priority? In *The ideal of equality*, ed. M. Clayton and A. Williams. London/New York: Palgrave Macmillan/St Martin's Press.

Pogge, T. 2003. Can the capability approach be justified? *Philosophical Topics* 30(2): 167–228.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 2001. *Justice as fairness*. Cambridge, MA: Harvard University Press.

Roemer, J.E. 1986. Equality of resources implies equality of welfare. *Quarterly Journal of Economics* 101: 751–784.
Roemer, J.E. 1998. *Equality of opportunity.* Cambridge, MA: Harvard University Press.
Sen, A. 1999. *Development as freedom*. New York: Knopf.
Temkin, L. 2002. Equality, priority, and the levelling down objection. In *The ideal of equality*, ed. M. Clayton and A. Williams. London/New York: Palgrave Macmillan/St Martin's Press.
Walzer, M. 1983. *Spheres of justice*. New York: Basic Books.
Williams, A. 2002. Dworkin on capability. *Ethics* 113: 23–39.

Egypt, Economy of

Barry Turner

Keywords
Foreign direct investment; International Monetary Fund; Mubarak; Piasters

JEL Classification
O53; R11

Overview

Egypt, one of the most economically diversified countries in the Middle East, has been embroiled in economic turmoil since the overthrow of Hosni Mubarak in 2011.

The 1980s was a decade of macroeconomic disorder in Egypt before an IMF-backed reform programme in the 1990s helped it achieve economic stability. However, the late 1990s saw a downturn and privatization efforts stalled in the early 2000s. In 2004 an economically liberal cabinet was appointed and the reform agenda was revived, with President Mubarak investing political capital in structural reforms to generate jobs and promote foreign investment.

Growth in Egypt averaged 6.4% per year between 2005 and 2008, underpinned by record levels of foreign direct investment and a favourable external environment. The economy held up relatively well during the global financial crisis, with a decline in remittances and external demand partially offset by resilient domestic demand and strong performances in the construction, communications and trade sectors.

In 2009 agriculture accounted for 13.7% of GDP, industry 37.3% and services 49.0%.

GDP fell by 9% in the first quarter of 2011 following the revolution in January of that year. Revenues from tourism collapsed, triggering a slide in foreign reserves. Unemployment climbed to 13% and gross public debt rose to nearly 100% of GDP by late June 2013. Long-term stability is likely to remain elusive as the country struggles through its political transition.

Currency

The monetary unit is the *Egyptian pound* (EGP) of 100 *piastres*. Inflation rates (based on IMF statistics) for fiscal years:

Faced with slowing economic activity, the country devalued the Egyptian pound four times in 2001. In January 2003 the Egyptian pound was allowed to float against the dollar after years of a government-controlled foreign exchange regime. In June 2009 foreign exchange reserves were US$29,278 m., gold reserves totalled 2.43 m. troy oz and total money supply was £E182,991 m.

Budget

The financial year runs from 1 July. Budgetary central government revenues in 2008–09 (provisional) were £E282,505 m. and expenditures £E308,070 m. Taxes on income, profits and capital gains accounted for 28.4% of revenue and taxes on goods and services 22.2%. Main items of expenditure were subsidies (30.5%) and compensation of employees (24.7%).

Performance

Real GDP growth rates (based on IMF statistics):

Total GDP in 2012 was US$262.8 bn.

2002	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
2.4%	3.2%	8.1%	8.8%	4.2%	11.0%	11.7%	16.2%	11.7%	11.1%

2002	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
3.2%	3.2%	4.1%	4.5%	6.8%	7.1%	7.2%	4.7%	5.1%	1.8%

Banking and Finance

The Central Bank of Egypt (founded 1960) is the central bank and bank of issue. The *Governor* is Hisham Ramez.

In 2003, four major public-sector commercial banks accounted for some 77% of all banking assets: the National Bank of Egypt (the largest bank, with assets of £E299 bn. in June 2010), the Banque Misr, the Bank of Alexandria and the Banque du Caïre. There were 40 banks in total in 2008. Foreign banks have only been allowed to operate since 1996.

Foreign direct investment inflows, which were just US$237 m. in 2003, rose to US$11.6 bn. in 2007, but declined to US$6.4 bn. in 2010.

In 2010 external debt totalled US$34,844 m., representing 16.2% of GNI.

There are stock exchanges in Cairo and Alexandria.

Central Bank of Egypt: http://www.cbe.org.eg
Bank of Alexandria: http://www.alexbank.com
Banque du Caire: http://www.bdc.com.eg
Banque Misr: http://www.banquemisr.com
National Bank of Egypt: http://www.nbe.com.eg/en/main.aspx

See Also

- ▶ Energy Economics
- ▶ International Monetary Fund
- ▶ Islamic Economic Institutions
- ▶ Islamic Finance
- ▶ Oil and the Macroeconomy
- ▶ Organization of the Petroleum Exporting Countries (OPEC)

E

Einaudi, Luigi (1874–1961)

F. Caffé

Keywords
Economic liberalism; Einaudi, L.; Mill, J. S.; Public finance; Tax incentives for saving; Taxable income

JEL Classifications
B31

An outstanding Italian economist and influential figure on the broader political and cultural scene, Einaudi was born in Carru (Piedmont) on 24 March 1874 and died in Rome on 30 October 1961. He graduated in law from Turin in 1895 and then, while continuing with this studies, embarked on a career in journalism. The success he achieved in both fields underlined his rare talent and his endless capacity for work. In fact, his academic progress was so rapid that in 1907 he was appointed as professor of public finance at the University of Turin. Meanwhile, he wrote articles for the most influential Italian daily newspaper of the period, the *Corriere delle Serra*, which not only brought him national recognition but also earned him the reputation of 'educator' of the entire country. He became a member of the Senate in 1919, but retired from all political and public activity with the advent of fascism. Towards the end of the First World War he went into exile in Switzerland. On his return, he was appointed Governor of the Bank of Italy (1945), Vice-

President of the Cabinet and Minister in charge of the Budget (1947), and was finally elected President of the Republic of Italy (1948–1955). At the end of his seven-year presidential term of office, he was made a life member of the Senate.

The most important aspect of Einaudi's achievements is the use he made of his academic and journalistic ability, as foundations for his activity as a statesman and politician. In addition, close study of his strictly scientific works reveals the extent to which he drew on the wealth of knowledge and experience which he had gained also in other fields. The 3,800 recorded items of Einaudi's works cover such a wide range of interests that it is necessary here to concentrate on his contributions to the study of public finance and his ideas on economic policy. Einaudi's main contributions to the study of public finance were investigations, based on the classical ideas of John Stuart Mill, which gave a solid logical basis to the principle of the exclusion of savings from taxable income; his research into the theory of capitalization of taxation; his critical and constructive contributions on the effects of certainty and stability of fiscal principles; his important analysis of the concept of taxable income which he identified with normal income, or, in other words, with the average income potentiality of the person subject to taxation.

Einaudi's position vis-à-vis public intervention in the economy was not hostile in principle, though he undoubtedly took a limited view of state interference in economic life. Since, for Einaudi, 'All liberties were jointly liable', autonomous sources of income were a necessity to prevent people from being subjected to a single centralizing order of the state. He asserted this during the 20 years of fascism, when he continued to teach with the same independence of mind and without compromising his fidelity to economic liberalism. Even though Einaudi had been stressing the usefulness of productive public expenditure since 1919, he showed a singular lack of comprehension of the Keynesian contribution, in the belief that it would be an inevitable cause of inflation.

Selected Works

On Luigi Einaudi himself there is a Bibliografia degli scritti edited by Luigi Firpo under the auspices of the Bank of Italy, Turin, 1971. It is useful to divide his work into the three main areas which he outlined: theory, politics and history. Representative works of the three sections are as follows:

1912. *Intorno al concetto di reddito imponibile e di un sistema di imposte sul reddito consumato*. Turin: V. Bona.

1919. *Osservazioni critiche intorno alla teoria dell'ammortamento dell'imposta e teoria delle variazioni nei redditi e nei valori capitali sussequenti all'imposta*. Turin: Fratelli Bocca.

1929. *Contributo all ricerca della 'ottima imposta'*. Milan: Bocconi.

1938. *Miti e paradossi delli giustizia tributaria*. Turin: Luigi Einaudi.

The following handbooks are available:

1914. *Corso di scienza delle finanze*. Turin: Tip. e Bono.

1932–66. *Principi di scienza delle finanze*. Turin: La Riforma Sociale.

1932. *Il sistema tributario italiano*. Turin: La Riforma Sociale.

With reference to the history of finance and the history of ideas see:

1908. *La finanza sabauda all'aprirsi del secolo XVIII e durante la guerra di successione spagnola*. Turin: Società Tip. Editrice Nazionale.

1927. *La guerra e il sistema tributario italiano*. Bari: Laterza.

1953. *Saggi bibliografici e storici intorno alle dottrine economiche*. Rome: Ediz. Storia e Litteratura.

Einaudi's journalistic work has been largely collected in eight volumes comprising the *Cronache economiche e politiche di un trentennio* (1893–1925), Turin: Ed. Einaudi, 1959–65, and in *Lo scrittoio del Presidente 1948–1955*, Turin: Ed. Einaudi, 1956. For many years Einaudi was Italian correspondent for the *Economist*.

Einzig, Paul (1897–1973)

Brendan Brown

Einzig was born in Brasov, Transylvania (Austria-Hungary). He was both a prolific, widely read author on international monetary topics and a renowned journalist. Educated in Hungary and France, he received his PhD from the University of Paris. In 1919, Einzig settled in the UK. Soon he became the Paris correspondent of the *Financial News* and was appointed its political editor in 1929. When the *Financial News* was bought by the *Financial Times,* Einzig became the political editor of the latter newspaper. Also he wrote the daily 'Lombard Street Column' during the mid- and late 1930s and many feature articles on currency questions. One of his top 'scoops' as a journalist was the revelation in 1943 of how the Swiss National Bank was buying looted gold from the Reichsbank on a huge scale. Already in 1939, Einzig's book *The Bloodless Invasion* had provided an original account of how Nazi Germany in its exchange rate policies exploited South-East Europe.

Einzig wrote more than 50 books on financial topics. Perhaps *A Dynamic Theory of Forward Exchange* (1961) is the best example of his powerful combination of economic, practical and historical knowledge. The book has a section describing the methods of intervention by central banks in forward exchange markets in the interwar period – and also by the Austrian and Russian central banks in the late 19th century. Einzig takes issue with the 'static theory of forward exchange' in which forward rates are shown as determined by given international interest rate differentials. He stresses that these themselves are influenced by speculation in the forward market. Einzig showed that except in the case of perfect arbitrage, forward markets have to be considered explicitly in an analysis of international short-term capital movements.

In *Primitive Money, in its Ethnological, Historical and Economic Aspects*, Einzig looks at how different commodities came to be used as money in primitive and ancient society. He refutes the hypothesis that money developed primarily through the progress of division of labour and the resulting complexity of trade, which made barter increasingly cumbersome. Much more important was the designation of a commodity for use in non-commercial payments (religious sacrifices, blood money, bride prices etc.).

Selected Works

1931. *Behind the scenes of international finance*. London: Macmillan.
1939. *The bloodless invasion.* London: Macmillan.
1940. *Europe in chains.* London: Penguin.
1949. *Money in its ethnological aspects*. London: Macmillan.
1954. *Monetary policy.* London: Penguin.
1960. *In the centre of things.* London: Hutchinson.
1961a. *Theory of foreign exchange.* London: Macmillan.
1961b. *A dynamic theory of forward exchange*. London: Macmillan.
1962. *A history of foreign exchange*. London: Macmillan.
1967. *Foreign exchange crisis.* London: Macmillan.

Eisner, Robert (1922–1998)

James K. Galbraith

Abstract

Robert Eisner, a leading American macroeconomist and theorist of the investment function, was an architect of the Keynesian ascendancy in post-war America. He developed the accounting foundations of Keynesian macroeconomics, finally producing a Total Income System of Accounts. His ideas found

application in his later, policy-oriented writings on the budget deficit, the current account, and the Social Security system. His embrace of capital budgeting underpinned a strong advocacy of liberal expenditure on infrastructure, education, and research and development. He was throughout motivated by a commitment to larger social goals, especially full employment, peace, and justice.

Keywords

American Economic Association; Budget deficits; Capital budgeting; China; One-child policy; Collective bargaining; Depreciation allowance; Economists Allied for Arms Reduction; Eisner, R.; Full employment; Invariant multiplier; Investment function; Keynesian Revolution; Liquidity preference; Liquidity trap; Natural rate of unemployment (NAIRU); Office of Price Administration; Peace economics; Permanent income hypothesis; Replacement costs; Social Security; Total Income System of Accounts; Unit relative price elasticity

JEL Classifications

B31

Robert Eisner, a leading American macroeconomist and theorist of the investment function, graduated in history from College of the City of New York in 1940, took an MA in sociology from Columbia University in 1942 and, following service in the army and the Office of Price Administration, a Ph.D. in economics under Fritz Machlup at Johns Hopkins University in 1951. He joined the faculty of Northwestern University in 1952, rising to hold the William R. Kenan Professorship of Economics from 1974 until his retirement in 1994. He served as President of the American Economic Association in 1988.

Eisner was an architect of the Keynesian ascendancy in post-war America. Much of his work was devoted to technical developments in that tradition; his singular distinction lay in taking the accounting foundations of Keynesian macroeconomics seriously and in developing their implications with utmost rigour. This thread runs through his writing from his earliest papers on the 'Invariant Multiplier', the permanent income hypothesis, liquidity preference and the liquidity trap. It reaches its apogee in his work on a Total Income System of Accounts (TISA). It suffuses his later, policy-oriented writings on the meaning and implications of deficits in the budget, current account, and Social Security system.

No shrinking violet, Eisner liked to call his shots. Thus, H. S.Houthaker 'has not performed [a] test correctly'; 'Bronfenbrenner and Mayer. . . confound. . . issues of elasticity with those of slope'; 'Re-estimation with Pifer's data and application of appropriate statistical tests contradict Pifer's conclusions' (1998a, pp. 8, 27, 48). The tone is ever tactful, the intent always the pursuit of truth, the subtext a certain delight in finding the exact, fatal weakness of an opposing view. Late in his life, this author heard Eisner speak to a room of senior officials in China on the error and futility of the one-child policy, a delicate issue which he raised in the same spirit and with deeply impressive effect.

Underpinning his technical precision lay an unflagging commitment to larger social goals, especially full employment, peace, and justice. Eisner actively advocated all three throughout his career, but especially in the later years when he appeared frequently on the opinion pages of the *Wall Street Journal,* as a leading director of Economists Allied for Arms Reduction, and in causes devoted to the advancement of women in the economics profession.

For instance, in a 1952 paper in the *American Economic Review* (1998a, 106–17) Eisner analysed the relationship of replacement costs to depreciation allowances in a growing economy. In doing so he called attention to the fact that growth in the latter usually exceeded that in the former, resulting in reported profits that were understated for purposes of both taxation and collective bargaining. Pointedly, he suggested the work ought to interest both revenue officers and trade unionists.

Yet Eisner's views were often unfashionable and politically inconvenient. In important papers in the 1980s, at a time when Democrats had taken the veil of fiscal virtue, he undertook with Paul

Pieper to show that (among numerous other difficulties with budget accounting) inflation had rendered the deficit meaningless, introducing vast inconsistencies between the nominal budget deficit and the change in the real public debt. Thus, the Reagan deficits were far smaller than normally supposed, while those of Carter were surpluses in real terms – likely to produce fiscal drag and so to bear partial responsibility for the stagnation of those years. Correctly accounting for inflation, Eisner argued, might have forestalled the new classical critique that led many in those years to abandon Keynesian principles.

A closely related cause was the misunderstanding of 'national saving' and the fallacious popular argument that to reduce deficits would lead to increased capital formation. In 1995, Eisner argued that to take the accounting relation between public and private saving

> as evidence that reducing the federal deficit must raise national saving should be recognized, on even the slightest reflection, as patently absurd. It is startlingly akin to the assumption, more than half a century ago, that saving and investment would be increased if we all undertook to save more by consuming less. Perhaps! But that is exactly the proposition to be proved, or supported by empirical evidence, not assumed. (1998a, p. 322)

Second only to correct reasoning, evidence mattered. In the 1990s Eisner took up arms against the 'governing myth' of economic policy, the natural rate of unemployment introduced by Friedman and Phelps in 1968. From this strangely selfdamaging justification for perpetually high unemployment, Eisner hoped for a 'NAIRU escape'. His method was largely econometric, and in what may have been his final paper, published in 1998 (1998a, pp. 454–87), he argued that a separate analysis of low-unemployment cases showed no relationship between full employment and rising inflation. This position was to be vindicated dramatically in the two years following his death.

Eisner embraced capital budgeting, so that the liabilities acquired by the government might be properly offset against corresponding assets. This position helped underpin a strong advocacy of liberal expenditure on infrastructure, education, and research and development. It also provides one bridge between the Keynesian Eisner and his counterpart, the theorist of investment, public finance, and peace economics and stalwart defender of Social Security, all of which he was.

Eisner's investigations of investment involved pioneering use of corporate records. They permitted cross-section analysis of firm decisions, showing that the concepts of macro models, such as the accelerator, operated differently on firms from different industries or with differing recent growth histories. In numerous studies, Eisner criticized neoclassical investment theories. Rejecting the notions of a desired capital stock and unit relative price elasticity, he adhered to a Keynesian relation of investment to expected profitability and of expected profits to the rate of growth. An important theme in this work concerns the appropriate level of aggregation at which to take measurements. Eisner found that firms appropriately assess the growth of their own industry to be the most relevant to profit prospects, not the inherently variable growth of individual firms or the potentially irrelevant growth of generalized aggregate demand.

Eisner's Total Income System of Accounts marked the peak of his campaign to rationalize economic measurement and theory. The importance of changing household relations appears vividly in his initial motivation for this work: 'What happens to income, output, and productivity when clotheswashing moves from the washtub and the professional laundry to the laundromat and to the automatic washer and dryer. . .?' (1998b, p. 188). Particularly noteworthy is capital accumulation by households in a country where transportation is provided mainly and increasingly by private car. The challenge of TISA remains to be taken up by most economists and national income statisticians.

Finally, midway through the Vietnam War Eisner deflated the view that President Johnson might have forestalled inflation by raising taxes; the only sure way to that end, he showed, would have been to avoid the war. This insight led to papers on the 'staggering cost' of the Vietnam war, much in the spirit of total accounts, and on post-cold war disarmament. Equally, to the end of his life Robert Eisner defended Social Security from all those who would cut it. Spurious and persistent allegations of financial 'crisis' notwithstanding, he

believed that a rich and civilized society can, and should, provide decent incomes and care for its old.

See Also

- Government Budget Constraint
- Labour Supply
- National Accounting, History Of
- Social Security in the United States
- War and Economics

Selected Works

1986. *How real is the federal deficit?* New York: Free Press.
1994. *The misunderstood economy: What counts and how to count it.* Cambridge, MA: Harvard Business School Press.
1997. *The great deficit scares: The federal budget, trade, and social security.* New York: The Century Foundation.
1998a. *The Keynesian revolution, then and now: The selected essays of Robert Eisner*, vol. 1. Cheltenham: Edward Elgar.
1998b. *Investment, national income and economic policy: The selected essays of Robert Eisner*, vol. 2. Cheltenham: Edward Elgar.

Elasticities Approach to the Balance of Payments

Murray C. Kemp

Keywords
Elasticities approach to the balance of payments; Excess demand

JEL Classifications
F32

The substance of a theory is independent of the manner in which it is dressed. In particular, it is a matter of style only whether or not formulae are expressed in terms of elasticities of demand and supply, or in terms of ordinary derivatives. To speak of an 'elasticities approach' to the balance of payments is therefore to speak no sense at all.

However, behind the nonsensical label there hides a coherent and distinctive theory of what determines the response of a country's balance of payments to parametric changes in its rate of exchange, that is, to changes in the terms on which its currency exchanges for other currencies. The theory goes back to a paper published by Charles Bickerdike (1920).

Consider a simplified world containing just two countries (the 'home' country and the 'foreign') and producing and trading just two commodities. Let R be the price of foreign currency in terms of home currency, let p_i be the home price of the ith commodity in terms of home currency (so that, in arbitrage equilibrium, $p_i^* = p_i/R$ is the foreign price of the commodity in terms of foreign currency), and let B be the home balance of trade in terms of foreign currency. Then, writing $z_i(p_i)$ and $z_i^*(p_i^*)$ as the home and foreign excess demands for the ith commodity, Bickerdike's model of the balance of payments reduces to the system of three equations

$$\begin{aligned} z_i(p_i) + z_i^*(p_i/R) &= 0 \ (i = 1, 2) \\ B &= -(1/R)[p_1 z_1(p_1) + p_2 z_2(p_2)] \end{aligned} \tag{1}$$

In this system the rate of exchange R is treated as a parameter and p_1, p_2 and B as variables to be determined. Differentiating (1) with respect to R, solving for $\mathrm{d}B$ and the $\mathrm{d}p_i$, and converting to elasticities, we obtain

$$\mathrm{d}B = \left\{ -p_2^* z_2^* \left[\frac{\eta_1^*(1+\eta_1)}{\eta_1^* - \eta_1} - \frac{\eta_2^*(1+\eta_2)}{\eta_2^* - \eta_2} \right] - B \right\} \frac{\mathrm{d}R}{R} \tag{2}$$

and

$$\frac{\mathrm{d}p_i}{p_i} = \frac{\eta_i^*}{\eta_i^* - \eta_i} \frac{\mathrm{d}R}{R}, \ i = 1, 2 \tag{3}$$

where $\eta_i \equiv (dz_i/dp_i)(p_i/z_i)$ and $\eta_i^* \equiv (dz_i^*/dp_i^*)(p_i^*/z_i^*)$. In the special case in which B is initially zero, (2) takes the simpler form

$$dB = -p_2^* z_2^* \left[\frac{\eta_1^*(1+\eta_1)}{\eta_1^* - \eta_1} - \frac{\eta_2^*(1+\eta_2)}{\eta_2^* - \eta_2}\right]\frac{dR}{R} \quad (4)$$

Equation (2) is often referred to as the Bickerdike–Robinson–Metzler formula: however, the role of Robinson (1947) and of Metzler (1949) was that of expositor only.

Suppose for concreteness that the home country exports the first commodity and imports the second, so that η_1 and η_2^* are export-supply elasticities and η_2 and η_1^* import-demand elasticities. Suppose further that all marginal propensities to buy are positive, so that η_1 and η_2^* are positive, η_2 and η_1^* negative. Then for the balance of payments to improve in response to devaluation it suffices that the sum of the two import demand elasticities exceed 1 in magnitude, that is, that the Marshall-Lerner condition be satisfied. Thus Eq. (4) can be rewritten as

$$dB = -p_2^* z_2^* \left[\frac{\eta_1\eta_2^*(1+\eta_1^*+\eta_2) - \eta_1^*\eta_2(1+\eta_1+\eta_2^*)}{(\eta_1 - \eta_1^*)(\eta_2 - \eta_2^*)}\right] \times \frac{dR}{R}$$

with all terms of known sign except $(1 + \eta_2^* + \eta_2)$. For a positive response of the balance of payments to devaluation it suffices also that the terms of trade improve, or at least that they not worsen. For changes in the terms of trade are indicated by changes in p_1/p_2 and, from Eq. (3),

$$\frac{d(p_1/p_2)}{p_1/p_2} = \frac{dp_1}{p_1} - \frac{dp_2}{p_2} = \left(\frac{\eta_1^*}{\eta_1^* - \eta_1} - \frac{\eta_2^*}{\eta_2^* - \eta_2}\right)\frac{dR}{R}$$

If this expression is non-negative then, from (4), dB must be positive.

Bickerdike's theory is very special in that the excess demand for each commodity depends on the money price of that commodity only. Implicitly, all 'cross' price elasticities are set equal to zero. For more general theories and, in particular, more general versions of (4), the reader is referred to Negishi (1968), Kemp (1970), Dornbusch (1975) and Kyle (1978).

Bibliography

Bickerdike, C.F. 1920. The instability of foreign exchange. *Economic Journal* 30: 118–122.

Dornbusch, R. 1975. Exchange rates and fiscal policy in a popular model of international trade. *American Economic Review* 65: 859–871.

Kemp, M.C. 1970. The balance of payments and the terms of trade in relation to financial controls. *Review of Economic Studies* 37: 25–31.

Kyle, J.F. 1978. Financial assets, non-traded goods and devaluation. *Review of Economic Studies* 45: 155–163.

Metzler, L.A. 1949. The theory of international trade. In *A survey of contemporary economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

Negishi, T. 1968. Approaches to the analysis of devaluation. *International Economic Review* 9: 218–227.

Robinson, J. 1947. *Essays in the theory of employment*. 2nd ed. Oxford: Basil Blackwell.

Elasticity

Peter Newman

Abstract

Formally invented by Marshall, the concept of elasticity of demand goes beyond the notion, which can be found in classical economics, that demand varies less or more than price. The crucial property that alone makes elasticity so important in pure and applied economics is that the elasticity measure is *invariant* to changes in units of measurement of quantities and prices.

Keywords

Arc elasticity; Cournot, A. A.; Dimensional analysis; Duality; Elasticity; Elasticity of demand; Elasticity of substitution; Elasticity of supply; Invariance to transformation; Mill, J. S.

JEL Classifications

D0

One day in the winter of 1881–2 Alfred Marshall came down from the sunny rooftop of his hotel in Palermo 'highly delighted', for he had just invented elasticity of demand (Keynes 1925, pp. 39 n. 3, 45 n. 2). So delighted was he that within a mere four years he had introduced the word *elasticity* into the technical literature of economics (Marshall 1885), which by his own standards was rushing pell-mell into print. But if the speed of its introduction was uncharacteristic the manner of it was not, tucked away as it was at the end of a lecture dull even for its time, and giving no hint that elasticity was new and exciting (1885, p. 187).

The notion that demand varies less or more than price can of course be found rather often in classical economics, especially in John Stuart Mill (Edgeworth 1894, p. 691). But to turn that trite idea into something useful requires a firm grip on the prior idea of quantity demanded *at a price*. So it is not surprising that the only ancient who came close to Marshall's idea was Cournot himself, the inventor of (among much else) the demand function.

In fact Cournot came so close that it is hard to understand, first, why he did not go all the way, and second, why Marshall gave him no credit for showing that way. Such lack of generosity is the more puzzling since we know that between the time when (according to Mrs Marshall) he invented elasticity, and the late spring of 1882 when he first drafted the chapter on Elasticity for the *Principles*, Marshall reread Cournot (Whitaker 1975, vol. 1, p. 85).

Starting with the demand function $D = F(p)$, Cournot pointed out that $pF(p)$ is total revenue, so that for maximum revenue the price p must be such that $F(p) + pF'(p) = 0$ (1838, p. 56). Thus total revenue will increase or decrease with increa se in price according as $\Delta D/\Delta p$ is larger or smaller than D/p, where ΔD is the absolute value of the change in quantity demanded.

> Commercial statistics should therefore be required to separate articles of high economic importance into two categories, according as their current prices are above or below the value which makes a maximum of $pF(p)$. We shall see that many economic problems have different solutions, according as the article in question belongs to one or other of these two categories. (Bacon's translation 1897, p. 54)

Let f be a real-valued nonzero differentiable function whose domain is some open interval I of the real line. In conformity with Marshall's Mathematical Appendix (1890, Note IV, pp. 738–40), the *elasticity of f at the point x,* denoted by $\eta_f(x)$, is defined here to be the *number* $xf'(x)/f(x)$. *The function* $\eta_f(x)$,defined by this formula is called the *elasticity of f.* To define the elasticity of *demand,* some authors prefer to follow the convention $f(x) = -xf'(x)/f(x)$, which is not used here. Unfortunately there is no standard notation for elasticity, since the obvious candidates are already taken, e for e and E for the expectations operator.

Cournot's critical value of p, his criterion for sorting out commodities, is simply that p^* for which $\eta_f(p^*) = -1$; he was close indeed. However, unlike Marshall (who is crystal clear on the point) there is no trace in Cournot of the crucial property that the elasticity measure is *invariant* to changes in units of measurement of quantities and prices, and it is this property alone that makes it so important in pure and applied economics.

A little calculus will prove such invariance, but is more enlightening to apply the dimensional analysis of Jevons and Wicksteed. Let the dimension of x be X and that of $f(x) = y$ be Y, so that $f'(x)$ has dimension YX^{-1}. The dimension of $\eta_f(x)$ is then $X \cdot YX^{-1} \cdot Y^{-1}$ and everything cancels. The elasticity of f at x is a pure number, unaffected by change in the units of either x or y. (This application is so obvious that the most plausible explanation of why it was not included in Wicksteed 1894, is that his entry was actually written before Marshall's *Principles* appeared.) Although invariance to transformation of units is the key property of elasticities, partly as a consequence the measure has a number of other agreeable properties. For example, it is easily seen that $\eta_f(x) = d \log f(x)/d \log x$, which paves the way for a whole calculus of elasticities in terms of logarithmic derivatives (Champernowne 1935; Allen 1938, pp. 251–4). One simple application of this calculus is the formula $\eta_{fg}(x) = \eta_f(x) + \eta_g(x)$, where fg is the product of f and g (with a corresponding formula for the quotient function f/g), while another is the characterization of constant elasticity functions as those which are linear in logarithms, that is, of Wicksell–Cobb–Douglas type. Incidentally,

Douglas's paper of 1927 was apparently intended to introduce elasticity of supply, which is odd since it had already appeared 20 years before (and rather late at that) in the fifth edition of the *Principles* (see Marshall 1961, vol. 2, p. 521).

The extension of elasticity to functions of more than one variable is easy – one simply uses the partial derivatives f_i rather than the derivative f' – and is staple fare in textbooks (see for example Allen 1938, pp. 310–12). However, many of those textbooks underplay another useful property of elasticities of strictly monotonic functions (such as the usual demand and supply curves) which follows from the inverse function theorem. Considering just functions of one variable, if we write $\Phi = f^{-1}$then from that theorem $\Phi' = f^{-1}$, so from this and the definition of elasticity,

$$\eta_\Phi(y) = y\Phi'(y)/\Phi(y) = f(x)/xf'(x) = \left(\eta_f(x)^{-1}\right),$$

that is, the elasticity of the inverse function is the inverse of the elasticity. Two obvious applications of this to the elementary theory of the firm are:

(i) Since the revenue function is $R(q) = pq = q\Phi(q)$,

$$\begin{aligned}\text{marginal revenue } (mr) &= \Phi(q) + q\Phi'(q)\\ &= \Phi(q)[1 + ?(q\Phi'(q)/\Phi(q)] = \Phi(q)\\ &\quad \times (1 + \eta_\Phi(q)),\end{aligned}$$

from which one can derive the more usual but less intuitive formula $mr = p[1 + (1/\eta_f(p))]$; and (ii) since at the firm's profit maximizing output marginal cost $mc = mr$, the Lerner (1934) measure of monopoly power (p –mc)/p may be written $[\Phi(q) - mr]/\Phi(q) = 1 - [\Phi(q)(1 + \eta_\Phi(q)]/\Phi(q) = -\eta_\Phi(q)$.

Arc elasticity, which is really ordinary elasticity with the index number problem thrown in, was introduced quite early by Dalton (1920, pp. 192–7). But the heyday of elasticities of all kinds came later, in the 1930s, so much so that it is small wonder that in the immediate post-war period Samuelson (1947, pp. 4–5) used elasticity statements to exemplify what he meant both by 'meaningful theorems' and by non-meaningful theorems in economics. A peculiar aspect of some of the elasticity measures introduced then was their definition not in terms of the properties of a given *function f* (as here), but rather as the ratio of proportionate change in one variable to proportionate change in another, allegedly causative, variable, without any explicit functional relationship intervening. Thus with Hicks's 'elasticity of expectations' (1939, p. 205) there *is* no 'expectation function' of which it is an elasticity, as that term is defined above. Similarly, although the elasticity of substitution (σ) invented by Hicks (1932) and Robinson (1933) immediately provoked many articles in response (for example, Lerner 1933), at no time was a 'substitution function' introduced whose elasticity it was. The lack of a generating function for σ might help to explain why its use often occasions technical difficulty.

It is of some interest to apply duality theory to the problem of deriving simple formulas for entities like σ (cf. Woodland 1982, p. 31). Consider the elasticity of substitution σbetween two consumer's goods x and y, with no restriction being placed on preferences apart from the smoothness conditions implicit at this level of analysis. First, take advantage of homogeneity in both the ordinary and compensated demand functions to write the former function as $f(p, m)$ and the latter as $h(p, t)$, where p is the price of x in terms of y, m is the consumer's income in terms of y, and t is the *maximized* level of utility for the price-income situation (p, m). Put $x^* = f(p, m)$. Finally, observe that σ is wholly determined by the price slope corresponding to p together with the indifference curve corresponding to t, so that we may write $\sigma = \sigma(p, t)$.

From a modern version of the fundamental equation of value theory (Hicks 1939, p. 309),

$$f_p(p, m) = h_p(p, t) - x^* f_m(p, m) \qquad (1)$$

where f_p, h_p and f_m are, in sequence, the partial derivatives of f and h with respect to p, and of f with respect to m. Multiplying (1) by $p/f(p, m)$ and writing $\eta_{fp}, \eta_{fm}$for the two partial elasticities of f, we obtain

$$\eta_{fp}(p,m) = ph_p(p,t)/x^* - px^* mf_m(p,m)/(mf(p,m)) = ph_p(p,t)/x^* - k\eta_{fm}(p,m) \quad (2)$$

where $k = px^*m$, that is, the fraction of m spent on x. Now since t is the maximized level of utility, given local non-satiation $x^* = h(p, t)$. Hence, the first term on the right-hand side of (2) is $\eta_{hp}(p, t)$, the partial elasticity of h with respect to p, and (2) becomes

$$\eta_{fp}(p,m) = \eta_{hp}(p,t) - k\eta_{fm}(p,m). \quad (3)$$

A standard result of Hicks and Allen (1934, see Hicks 1981, p. 20) for the two-good case can be written in the present notation as

$$-\eta_{fp}(p,m) = k\eta_{fh}(p,m) + (1-k)\sigma(p,t) \quad (4)$$

so from (3) and (4),

$$(k-1)\sigma(p,t) = \eta_{hp}. \quad (5)$$

Let the cost (expenditure) function for this problem be $c(p, t)$, and denote its partial derivative with respect to p by c_p. Then, writing η_{cpp} for the partial elasticity of c_p with respect to p, since Shephard's Lemma implies $c_p = h$ we have

$$\eta_{hp}(p,t) = \eta_{cpp}(p,t). \quad (6)$$

Now $k = px^*/m = ph(p, m)/m = pc_p(p, t)/m$. Because t is the maximized level of utility $m = c(p, t)$, so $k = pc_p(p, t)/c(p, t) = \eta_{cp}(p, t)$, where η_{cp} is the partial elasticity of c with respect to p. Substituting from this and (6) into (5),

$$\sigma(p,t) = \eta_{cpp}(p,t)/\big(\eta_{cp}(p,t) - 1\big). \quad (7)$$

Thus the elasticity of substitution in this two-good case can be expressed entirely in terms of the cost function.

See Also

▶ Marshall, Alfred (1842–1924)

Bibliography

Allen, R.G.D. 1938. *Mathematical analysis for economists*. London: Macmillan.

Champernowne, D.G. 1935. A mathematical note on substitution. *Economic Journal* 15: 246–258.

Cournot, A.A. 1838. *Recherches sur les principes mathematiques de la theorie des richesses.* Paris: Hachette. New edn, ed. G. Lutfalla, Paris: Riviere, 1938. English trans. by N.T. Bacon, 1897. Reprinted, New York: A.M. Kelley, 1960.

Dalton, H. 1920. *Some aspects of the inequality of incomes in modern communities*. London: Routledge.

Douglas, P.H. 1927. Elasticity of supply as a determinant of distribution. In *Economic essays contributed in honor of John Bates Clark*, ed. J.H. Hollander. New York: Macmillan.

Edgeworth, F.Y. 1894. Elasticity. In *Dictionary of political economy*, ed. R.H.I. Palgrave, vol. 1. London: Macmillan.

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

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

Hicks, J.R. 1981. *Collected essays on economic theory.* Vol. 1: Wealth and welfare. Cambridge, MA: Harvard University Press.

Hicks, J.R., and R.G.D. Allen. 1934. A reconsideration of the theory of value. *Economica* 1: 52–76, 196–219. Reprinted in Hicks (1981, vol. 1).

Keynes, J.M. 1925. Alfred Marshall, 1842–1924. In Pigou (1925).

Lerner, A.P. 1933. The diagrammatical representation of elasticity of substitution. *Review of Economic Studies.* Reprinted in Lerner (1953).

Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1(June): 157–175. Reprinted in Lerner (1953).

Lerner, A.P. 1953. *Essays in economic analysis*. London: Macmillan.

Marshall, A. 1885. The graphic method of statistics. *Journal of the Royal Statistical Society. Reprinted in Pigou* 1925.

Marshall, A. 1890. *Principles of economics*. Vol. 1. London: Macmillan.

Marshall, A. 1961. *Principles of economics*. 9th (variorum) edn, with annotations by C.W. Guillebaud, 2 vols. London: Macmillan.

Pigou, A.C., ed. 1925. *Memorials of Alfred Marshall.* London: Macmillan. Reprinted, New York: A.M. Kelley, 1966.

Robinson, J.V. 1933. *The economics of imperfect competition*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of economics analysis*. Cambridge, MA: Harvard University Press.

Whitaker, J.K., ed. 1975. *The early economic writings of Alfred Marshall, 1867–1890,* 2 vols. New York: Free Press.

Wicksteed, P.H. 1894. Dimensions of economic quantities. In *Dictionary of political economy,* vol. I, ed. R.H.I. Palgrave. London: Macmillan.

Woodland, A.D. 1982. *International trade and resource allocation*. Amsterdam: North-Holland.

Elasticity of Intertemporal Substitution

Christopher Bliss

Abstract

The elasticity of intertemporal substitution (EIS) measures the willingness on the part of the consumer to substitute future consumption for present consumption. It plays a key role in the theory of consumption and saving, in particular in the life-cycle version of that theory.

Keywords

Consumption smoothing; Consumption theory; Elasticity of intertemporal substitution; Elasticity of substitution; Multiple equilibria; Overlapping generations models; Risk aversion; Savings

JEL Classifications

D0

The EIS and Consumption Theory

The elasticity of intertemporal substitution (EIS) is an important number in macroeconomic theory. It measures the willingness on the part of the consumer to substitute future consumption for present consumption. This parameter plays a key role in the theory of consumption and saving, in particular in the life-cycle version of that theory. For a start we examine the role of the EIS in a basic life-cycle model. In that model there is complete certainty concerning prices, future income, and preferences present and future. The consumer can lend and borrow at will at a single invariant rate of interest, subject only to a lifetime budget constraint. Preferences are additively separable. The consumer chooses present and future consumption to maximize:

$$\sum_{t=1}^{T} \delta^{t-1} U[c_t] \tag{1}$$

where c_t is consumption in period t, and $0 < \delta < 1$, and where δ is the rate at which utility is discounted. Lifetime utility (1) is maximized subject to the lifetime budget constraint:

$$\sum_{t=1}^{T} c_t \left(\frac{1}{1+r}\right)^{t-1} \leq \sum_{t=1}^{T} y_t \left(\frac{1}{1+r}\right)^{t-1} \tag{2}$$

where the y values are incomes in the various periods, and r is the real rate of interest. Assuming positive consumptions in all periods, the maximization of (1) requires:

$$\frac{dU[c_t]}{dc_t}\delta^{t-1} - \lambda\left(\frac{1}{1+r}\right)^{t-1} = 0 \tag{3}$$

where λ is the Lagrange multiplier. From (3), taking logs:

$$\ln\frac{dU[c_{t+1}]}{dc_{t+1}} - \ln\frac{dU[c_t]}{dc_t} = \ln\left(\frac{1}{1+r}\right) - \ln\ \delta \tag{4}$$

Differentiating (4) with respect to r and holding c_t constant gives:

$$\frac{\dfrac{d^2U[c_{t+1}]}{dc^2{}_{t+1}}}{\dfrac{dU[c_{t+1}]}{dc_{t+1}}}\frac{dc_{t+1}}{dr} = -\frac{1}{(1+r)} \tag{5}$$

A useful way of writing (5) is:

$$\frac{1}{c_{t+1}}\frac{dc_{t+1}}{dr} = -\frac{1}{(1+r)}\frac{\dfrac{dU[c_{t+1}]}{dc_{t+1}}}{\dfrac{d^2U[c_{t+1}]}{dc_{t+1}}} \tag{6}$$

Or equivalently:

$$\frac{1}{c_{t+1}}\frac{dc_{t+1}}{dr} = \frac{\sigma(c_{t+1})}{1+r} \tag{7}$$

where σ is the EIS, defined as:

$$\sigma(c) = -\frac{\dfrac{dU[c]}{dc}}{c\dfrac{d^2U[c]}{dc^2}} \tag{8}$$

Equation (7) indicates that the size of the EIS will be a crucial determinant of how far consumption levels will respond to changes in the interest rate.

The effect of a small change in r analysed above is a standard partial equilibrium result, in which enough is held constant to obtain a definite result. The calculation shows how two solution paths compare with regard to c_{t+1}, as r is varied slightly, when for each of these paths c_t takes the same optimal value. For that special case, (7) says that c_{t+1} increases with r, which is to say that c_{t+1} increases relative to c_t. In that particular sense a small increase in r encourages saving. Even for the two-period model popular for classroom exposition, it cannot be shown that a rise in r encourages saving. However in the two-period model it is true for any separable lifetime utility function, as (1), that c_1 declines as r increases, provided that $c_2 > y_2$, the usual case. When r increases the substitution effect always favours lower early consumption. When $y_2 > c_2$, however, the income effect opposes the substitution effect, and the outcome is uncertain.

Equation (8) shows that the second derivative of the utility function, how curvy it is if one likes, is crucial in giving a specific value to the EIS. If:

$$U(c) = c^{\frac{\sigma-1}{\sigma}} \tag{9}$$

then the EIS is constant, independent of c, and equal to σ.

Consumption Smoothing and Risk Aversion

The EIS as defined in (8) is the same as the Arrow–Pratt measure of relative risk aversion. It is no accident that consumption substitution through time, with no uncertainty whatsoever, and risk aversion, where uncertainty is necessarily involved, should involve the same parameter. Absolute risk aversion is related to the willingness of a consumer to accept a lottery ticket in preference to a sum of money available for certain, the certain sum being lower than the expected value of the lottery. One can think of the extra expected value in the better-than-fair lottery as a premium needed to entice the agent to accept the risk. The higher is relative risk aversion, the larger must be the expected-value premium in the lottery. Arrow (1971, ch. 3) provides a detailed discussion, and references the parallel and independent work of Pratt.

Now consider the life-cycle maximization of (1) subject to (2). To make the explanation as simple as possible let δ and r both be zero. The consumer maximizes:

$$\sum_{t=1}^{T} U[c_t] \tag{10}$$

subject to the lifetime budget constraint:

$$\sum_{t=1}^{T} c_t \le \sum_{t=1}^{T} y_t \tag{11}$$

With U [ ] a concave function, it is evident that the consumer will consume at the same level in each period:

$$c_t = \frac{\sum_{t=1}^{T} y_t}{T} \tag{12}$$

In the particular sense defined by this special case, the consumer is averse to consumption variability over time. It is the same as the risk-averse consumer disliking variations in wealth when different states of the world are realized. That each period of time will certainly arrive, whereas only one state of the world will be realized, is irrelevant in the *ex ante* view of the consumer facing uncertainty. A risk-averse agent can be induced to accept a gamble if the odds are sufficiently favourable, that is, if the expected-value premium

is sufficiently large. Similarly, a life-cycle planner will opt for a non-constant consumption plan if it provides a larger total consumption sufficient to compensate for the unattractive variability. A positive rate of interest plays the same role as an expected-value premium. It is the sweetener that persuades the consumer to accept variability. For this reason it is no surprise to find that the extent to which the consumer will respond to the sweetener, in either case, is governed by precisely how much the consumer dislikes variability. And the EIS, or the coefficient of relative risk aversion, as the case may be, measures that dislike of variability.

The argument just completed ignores the part played by δ, the utility discount rate. The presence of a positive δ means that, were r zero, the consumer would choose a plan with consumption falling through time. Then a positive r, and especially an r greater than δ, persuades the consumer to select a consumption plan with consumption falling less rapidly or rising through time. How far an optimal plan responds to a given change in r is governed again by the EIS.

A Constant or a Variable Coefficient?

The EIS has been compared above to the coefficient of relative risk aversion. In the theory of risk aversion the emphasis is on the variability of the coefficient. On this turns the issue of whether the wealthy will be more or less willing to undertake risk than the poor. With the EIS the most common assumption is that it is a constant. A popular special case of (1) is:

$$U[c_1, c_2, \ldots, c_n] = c_1^{\frac{\sigma-1}{\sigma}} + \delta c_1^{\frac{\sigma-1}{\sigma}} + \ldots + \delta^{n-1} c_n^{\frac{\sigma-1}{\sigma}} \qquad (13)$$

This is the love-of-variety utility function of Dixit and Stiglitz (1977), with discounting added. The EIS measured at any of the consumptions above is σ.

The elegance and convenience of forms such as (13) has made them appealing. Thus Barro and Sala-i-Martin (1995), in their influential study of economic growth, assume that different countries or regions solve independent Ramsey optimal model problems. This leads to the condition:

$$\frac{1}{c}\frac{dc}{dt} = \sigma[AF_1\{k, 1\} - \delta] \qquad (14)$$

where F_1 is the marginal product of capital, c is consumption, k is capital, δ is the utility discount rate, A measures total factor productivity as it is affected by policy, culture, corruption, and so on, and σ is the EIS. The lower is k the larger is F_1. If this effect is not offset by poor countries having lower total factor productivities, and if all countries share the same values of δ and σ, then conditional β- convergence follows from (14), meaning that poor countries grow faster.

The poor will be reluctant to save if their value of σ is low. And this is a most plausible specification. When all the meals that one eats are small, it is rationally more difficult to postpone eating now for a larger meal later. This point has been recognized in the literature. For example, King and Rebelo (1993) allow for a utility function of the Stone–Geary form, where the consumer gives priority to a fixed basket of essentials until that basket has reaches a critical scale. With those preferences, the poorest consumers will not save at all, and there is the possibility of a poverty trap. The Stone–Geary utility function implies a zero value for the EIS at low consumptions, and positive values for higher consumptions.

The EIS in Consumption Studies

Many applied economists used to take the view that the value of σ is close to zero (see Hall 1988; Mankiw et al. 1985). This reflects the failure of consumption studies to find a significant effect of the rate of interest on saving. Such estimates are seriously biased if the consumer is constrained from borrowing freely (a feature ignored in the computations above) or if, as in Deaton (1992), most consumers save only to replenish

precautionary balances following negative shocks. Then the optimizing substitution-based theory does not apply. Blundell et al. (1994) and Attanasio and Browning (1995) show that representative consumer models give seriously misleading results when applied to aggregate consumption data. They use UK household expenditure data to model consumption at the individual level and obtain a greatly improved fit when they allow the rich to have a higher EIS than the poor. Does that mean that as economies grow richer over time, the average EIS will increase? This remains an unanswered question.

VEIS Functions

Let the utility function be chosen from a class of which the simplest case is:

$$U[c] = \int_0^c \exp\left\{\frac{1}{\beta x}\right\} dx \tag{15}$$

where β is a positive constant and c is the level of consumption. This is a VEIS utility function, where VEIS stands for *variable elasticity of intertemporal substitution*. Then:

$$\frac{dU[c]}{dc} = \exp\left\{\frac{1}{\beta x}\right\} > 0 \tag{16}$$

and:

$$\frac{d^2U[c]}{dc^2} = -\exp\left\{\frac{1}{\beta c}\right\}\frac{1}{\beta c^2} > 0 \tag{17}$$

U [•] is an increasing concave function. Now the EIS may be computed as:

$$-\frac{\dfrac{dU[c]}{dc}}{c\dfrac{d^2U[c]}{dc^2}} = \beta c \tag{18}$$

This increases linearly with consumption at rate β. The poor have a lower EIS and β- convergence will not necessarily prevail.

A Variable EIS in the Diamond Capital Model

In their deep study of the Diamond overlapping generations model with capital, De La Croix and Michel (2002) more or less dismiss the importance of multiple stable equilibria. To summarize, it is possible to obtain multiple stable steady-state solutions with simple functional forms, but these cases are unsatisfactory at best. If the production function is Cobb–Douglas and with a simple separable utility function, there are no cases of multiple stable steady states. With a logarithmic utility function and the constant elasticity of substitution in production $\rho > 0$, there can be two positive steady-states, but it may be that only the corner degenerate outcome is stable.

Rather than using given simple functional forms and looking for a few steady-state solutions, try for a continuum of solutions as follows. Assume:

$$-\frac{\dfrac{dU[c]}{dc}}{c\dfrac{d^2U[c]}{dc^2}} = \sigma(c) \tag{19}$$

where $\sigma(c)$ is an arbitrary positive increasing function of c. Then:

$$\frac{\dfrac{d^2U[c]}{dc^2}}{\dfrac{dU[c]}{dc}} = -\frac{1}{\sigma(c)c} \tag{20}$$

Integrating (20) gives:

$$\ln\frac{dU[c]}{dc} = -\int_a^c \frac{1}{\sigma(x)x}dx + \ln\ D \tag{21}$$

where a is a positive constant, and D is a constant of integration.

In a steady state solution to the Diamond model we must have:

$$\ln\frac{dU[c_1]}{dc} - \ln\frac{dU[c_2]}{dc} = \ln\ \delta - \ln\ R \tag{22}$$

where c_1 and c_2 are consumption in respectively the first and second period of a life, R is the gross

rate of return to saving, and δ is the discount factor. From (21) and (22):

$$\int_{c_1}^{c_2} \frac{1}{\sigma(x)x} dx = \ln\ \delta - \ln\ R \tag{23}$$

Now in steady state c_1, c_2 and R all depend upon capital per head k. If over some range of values of k every value gives a steady state, then (23) will be an identity in k. Let the per capita production function be Cobb–Douglas with coefficient α. Then (23) takes the form:

$$\int_{(1-\alpha)k^{\alpha}-k}^{k+\alpha k^{\alpha}} \frac{1}{\sigma(x)x} dx = \ln\ \delta - \ln\ \left(1 + \alpha k^{\alpha-1}\right) \tag{24}$$

When (24) is an identity in k, over an interval at least, then differentiating both sides of (24) gives:

$$\begin{aligned} &\frac{1}{\sigma(k+\alpha k^{\alpha})} \frac{1}{k+\alpha k^{\alpha}} \frac{1}{\sigma((1-\alpha)k^{\alpha}-k)} \\ &\quad \times \frac{1}{(1-\alpha)k^{\alpha}-k} \\ &\quad = \frac{\alpha(1-\alpha)k^{\alpha-2}}{1+\alpha k^{\alpha-1}} \end{aligned} \tag{25}$$

Take a given a value of k, and let $\sigma\ (c_1)$ values be known for the c_1 value implied by that k all the way up to the c_2 defined by the same k. Then $\sigma\ (c_2)$ values are determined by (25), which rolls out a solution for σ such that all values of k on a connected interval are steady-state equilibrium levels. The contrast to the case advanced by De La Croix and Michel is striking.

Concluding Remarks

The EIS is an important value, just as is its cousin, the coefficient of relative risk aversion. The use of a simple functional form has too often frozen the EIS as a constant. When it is allowed to vary, the β-convergence of growth theory is no longer secure; cross-section consumption studies perform better; and multiple equilibrium in the Diamond capital model is seen to be far more probable than previous studies indicate.

See Also

- ▶ Consumer Expenditure
- ▶ Consumer Expenditure (New Developments and the State of Research)

Bibliography

Arrow, K.J. 1971. *Essays in the theory of risk bearing*. Amsterdam: North-Holland.

Attanasio, O.P., and M. Browning. 1995. Consumption over the life-cycle and over the business cycle. *American Economic Review* 85: 1118–1137.

Barro, R.J., and X. Sala-i-Martin. 1995. *Economic growth*. New York: McGraw-Hill.

Blundell, R., M. Browning, and C. Meghir. 1994. Consumer demand and the life-cycle allocation for household expenditures. *Review of Economic Studies* 61: 57–80.

Deaton, A. 1992. Understanding consumption. In *Clarendon lectures in economics*. Oxford: Clarendon Press.

De La Croix, D., and P. Michel. 2002. *A theory of economic growth: Dynamics and policy in overlapping generations*. Cambridge: Cambridge University Press.

Dixit, A.K., and J.E. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Hall, R.E. 1988. Intertemporal substitution in consumption. *Journal of Political Economy* 96: 339–357.

King, R.T., and S.T. Rebelo. 1993. Transitional dynamics and economic growth in the neoclassical model. *American Economic Review* 83: 908–931.

Mankiw, N.G., J.J. Rotenberg, and L. Summers. 1985. Intertemporal substitution in macroeconomics. *Quarterly Journal of Economics* 100: 225–281.

Elasticity of Substitution

D. R. Helm

Keywords

CES production functions; Cobb–Douglas functions; Elasticity; Elasticity of substitution; Hicks, J. R.; Income effects; Indifference curves; Production functions; Robinson, J. V.; Substitution; Substitution effects; Value theory

JEL Classifications
D0

The concept of the elasticity of substitution, developed by Joan Robinson and John Hicks separately in the 1930s, represented an important addition to the marginal theory of the 1870s, in the tradition of Marshall, Edgeworth and Pareto. It brought together two concepts which were already well established in the literature – the ideas of elasticities (which derive from Mill) and those of substitution (which go back to Smith). The relationship defined by the concept is a mathematical one relating to utility and production functions, with considerable economic implications. It has two applications: to the theory of production, and in particular the isoquant relationship between factor inputs, and to consumer behaviour and the indifference curve. Let us look at each in turn.

The two inventors of the concept – Joan Robinson, in her *Economics of Imperfect Competition* (1933), and John Hicks in his *Theory of Wages* (1932) – each developed Marshall's formula for the elasticity of derived demand. Each defined the concept somewhat differently. For Hicks, the definition was the percentage change in the relative amount of the factors employed resulting from a given percentage change in the relative marginal products or relative prices, that is (following Samuelson 1968):

$$\sigma = \sigma_{12} = (F_1F_2/FF_{12})\sigma_{21},$$

where $F(V_1, V_2)$ is a standard neoclassical production function, and the subscripts are the partial derivatives. This is sometimes called the direct elasticity of substitution.For Joan Robinson, on the other hand, concerned with relative shares and hence distributional issues, the elasticity of substitution was defined as 'the proportionate change in the ratio of the amounts of the factors employed divided by the proportionate change in the ratio of their prices' (1933, p. 256):

$$\sigma = -\frac{\partial(V_1/V_2)/(V_1/V_2)}{\partial(W_1/W_2)/(W_1/W_2)}$$

where W_1 is the price of the V_1 factor.

These two definitions of the concept gave rise to a considerable debate in the early issues of the *Review of Economic Studies*, with in particular a notable contribution from Kahn (1933) concerned to identify how these concepts related to each other. It turns out that these two original definitions are identical when the production function is confined to two factors of production, where the partial derivatives of the production function are the marginal productivities of the factor inputs and yield the relevant factor prices. In addition, the contributors to the debate attempted to identify the implications of these somewhat abstract concepts. Amongst these were the joint determination by the elasticity of substitution and the factor supplies of the relative shares of the factor reward (wages and profits), and implications for the definition of imperfect competition with increasing returns to scale.

It is not surprising that it is with the cases where the restrictive neoclassical assumptions for the production function are not met that most interest arises. Two important developments are where production function involves three or more factors and in extending from Cobb–Douglas to constant elasticity of substitution (CES) production functions. But although considerable emphasis has been placed on the elasticity of substitution in production, it remains a technical concept concerning factor substitutability. It has no direct allocation consequence. Diminishing elasticity of substitution does not imply diminishing returns to scale, since for returns we must have prices. Thus it is restricted to describing the technical conditions of production. But, being a technical concept, it can be generalized to all forms of transformation.

Thus, as we noted above, along with a number of other concepts, these tools developed for production were taken over to consumer theory. Because of the implications the concept had for the development of consumer behaviour, and because of the insight which the resulting difficulties threw up concerning the concept more generally, this application is of special interest.

It was Hicks (and Allen) who made that step. While Joan Robinson's development of the concept was closely related to her extension of

Marshall's theory of the industry, Hicks was familiar with a very different approach to value theory, that of Edgeworth, Pareto and Walras. While Joan Robinson had focused on production substitutions, and hence isoquants, Hicks took the idea developed in that domain, and translated it across to consumer theory, and to the indifference curves which he had got from Edgeworth. In the two goods case, price elasticity could be represented in terms of his fundamental formula, according to which:

$$\text{Price elasticity} = k(\text{income elasticity}) + (1-k)(\text{e.s.})$$

where k is the total expenditure that is spent on the commodity. Thus, with income elasticity, consumer theory led into a representation of the effect of a price change in terms of the income and substitution effects, with elasticity being thus of prime importance in classifying goods by their demand characteristics.

But whereas the elasticity concept in production theory naturally led on to the possibility of measurement, that step in consumer theory was more contentious. For although this technical concept represented one important step in the development of the marginalist approach to the theory of value, the theory of demand behaviour requires a behavioural theory of choice. The elasticity of substitution with respect to the indifference curve is one technical component. But, as with production theory, prices, and in this case the budget line, are also required.

Technical concepts thus aided the formulation of modern consumer theory as outlined in Hicks and Allen's 'A Reconsideration of the Theory of Value' (1934) and the opening chapters of *Value and Capital* (1939), a path from which it has scarcely deviated. But, despite the mathematical elegance of this construction, it may be argued that it disguised many of the important underlying questions. The increased power of the indifference curve analysis begged the question of whether consumer preferences could in reality be represented in this abstract way. Ultimately, whether consumer behaviour is well described by concepts like the elasticity of substitution, depends upon whether preferences can be represented by complete, transitive, utility functions. Much recent evidence from psychologists and decision theorists suggests otherwise. Likewise for production theory, the concepts of capital and labour may be themselves ambiguous.

See Also

- ▶ CES Production Function
- ▶ Cobb–Douglas Functions
- ▶ Production Functions

Bibliography

Hicks, J.R. 1932. *The theory of wages*. London: Macmillan.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Hicks, J.R. 1970. Elasticity of substitution again: Substitutes and complements. *Oxford Economic Papers* 22: 289–296.
Hicks, J.R., and R.G.D. Allen. 1934. A reconsideration of the theory of value I–II. *Economica* 1 (Pt 1): 52–76. Pt II, May, 196–219.
Kahn, R.F. 1933. The elasticity of substitution and the relative share of a factor. *Review of Economic Studies* 1: 72–78.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Samuelson, P.A. 1968. Two generalizations of the elasticity of substitution. In *Value, capital and growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

Electricity Markets

James Bushnell and Catherine Wolfram

Keywords

Cost-based regulation; Electricity industry; Electricity markets; Experimental economics; Game theory; Incentive regulation; Investment; Liberalization; Market power; Natural monopoly; Oligopoly simulation analysis; Privatization; Real-Time pricing; State-Owned utilities

JEL Classifications
L66

In many parts of the world buyers and sellers now trade electrical energy in liberalized markets. These markets have partially replaced cost-based regulation and government ownership.

Since the 1980s, governments in many countries have privatized and restructured their electricity industries. Liberalized electricity markets now operate in much of Europe, North and South America, New Zealand and Australia. These changes were primarily motivated by the perception that the previous regimes of either state ownership or cost-of-service regulation yielded inefficient operations and poor investment decisions. Liberalization of the electricity industry also reflected the progression of a deregulation movement that had already transformed infrastructure industries, including water, communications and transportation, in many countries. Although electricity shares many characteristics with other deregulated industries, the differences have proven to be more important than the similarities. Electricity has been one of the most challenging industries to liberalize and in most places new layers of regulations have replaced the old.

Historically, electricity was viewed as a natural monopoly. Typically, a single utility company generated, transmitted and distributed all electricity in its service territory. In much of the world, the monopoly was a state-owned utility. Within the United States, private investor-owned companies supplied the majority of customers, although federally and municipally owned companies played an important minority role. These companies operated under multiple layers of local, state and federal regulation.

Restructured electricity markets share a common basic organization. The three segments – generation, transmission, and distribution – have been unbundled. Wholesale generation, no longer viewed as a natural monopoly, is priced through a market process. Transmission and distribution remain regulated, although in many cases some form of incentive regulation has replaced cost-of-service regulation or state ownership.

Most wholesale electricity is traded through long-term (a week or longer) forward contracts. Many markets also feature day-ahead auction-based exchanges. Because supply and demand must be continually balanced to preserve transmission stability, transmission system operators run real-time balancing markets. Prices in these high-frequency markets can be highly volatile since electricity is non-storable and real-time demand fluctuates dramatically. To meet unforeseen contingencies, transmission system operators also contract for and occasionally use standby or reserve generation services. Many markets reflect price differences across geographical locations when parts of the transmission grid are congested (Schweppe et al. 1988; Chao and Peck 1992). Game theorists and experimental economists are involved in the ongoing process of designing electricity markets (Wilson 2002), while empirical researchers have used detailed auction data to estimate how well predictions from theoretical models describe firm behaviour (Wolak 2000; Hortascu and Puller 2004).

At the retail level, the vision of liberalization was to provide customers a choice among competing retailers who would operate as either resellers or integrated providers with access to customers through a regulated common-carriage distribution network. In most restructured US markets, retail competition for residential customers is very weak (Joskow 2005). Retail competition is more advanced in the United Kingdom, although evidence suggests that customers have been slow to take advantage of the ability to switch to a lower-priced retailer (Waddams 2004). Several authors have noted the economic benefits of allowing retail prices to vary to reflect real-time changes in the wholesale prices, although this sort of real-time retail pricing has been slow to take hold in practice (Borenstein and Holland 2005; Joskow and Tirole 2004).

Oligopoly simulation analysis indicates the potential for serious market power problems because suppliers face extremely inelastic demand and entry requires long lead times (Green and Newbery 1992). Empirical work has indicated that market power has indeed been present, although to varying degrees in different

markets. Wolfram (1999) found that prices in England and Wales were lower than static oligopoly models would suggest. By contrast, extreme levels of market power in California contributed to record high prices in 2000–1 (Borenstein et al. 2002). The explanations for these differences have focused on variations in the threat of future regulation and in the extent of long-term fixed price contracts (Bushnell et al. 2005).

Although the main motivation for market liberalization was to improve economic efficiency, there have been few attempts to measure efficiency changes. Newbery and Pollitt (1997) and Fabrizio et al. (2004) find modest positive effects of market liberalization on, respectively, industry efficiency in the United Kingdom and plant-level efficiency in the United States.

As electricity industry restructuring moves forward, the major unresolved question is the degree to which public policy will influence investment decisions. Electric generating plants are long-lived, so while operating efficiency gains appear to be real, the potential gains from improved investment stand to be larger. Also, policies to limit the environmental impact of electricity generation could affect the types of technologies in which we invest.

See Also

- ▶ Competition
- ▶ Energy Economics
- ▶ Privatization Impacts in Transition Economies

Bibliography

Borenstein, S., and S. Holland. 2005. On the efficiency of competitive electricity markets with time-invariant retail prices. *RAND Journal of Economics* 36: 469–493.

Borenstein, S., J. Bushnell, and F. Wolak. 2002. Measuring market inefficiencies in California's deregulated wholesale electricity market. *American Economic Review* 92: 1376–1405.

Bushnell, J., E. Mansur, and C. Saravia. 2005. Vertical arrangements, market structure, and competition: An analysis of restructured US electricity markets. CSEM working paper WP-126. Berkeley: University of California Energy Institute.

Chao, H.P., and S. Peck. 1992. A market mechanism for electric power transmission. *Journal of Regulatory Economics* 10: 25–60.

Fabrizio, K., N. Rose, and C. Wolfram. 2004. Does competition reduce costs? Assessing the impact of regulatory restructuring on US electric generation efficiency. CSEM working paper WP-135. Berkeley: University of California Energy Institute.

Green, R., and D. Newbery. 1992. Competition in the British electricity spot market. *Journal of Political Economy* 100: 929–953.

Hortascu, A., and S. Puller. 2004. Understanding strategic bidding in restructured electricity markets: A case study of ERCOT. Working paper. Chicago: University of Chicago. Online. Available at http://econweb.tamu.edu/puller/AcadDocs/Hortacsu_Puller.pdf. Accessed 20 July 2005.

Joskow, P. 2005. The difficult transition to competitive electricity markets in the US. In *Electricity restructuring: Choices and challenges*, ed. J. Griffen and S. Puller. Chicago: University of Chicago Press.

Joskow, P., and J. Tirole. 2004. Retail electricity competition. CMI working paper 44. Cambridge, MA: MIT.

Newbery, D., and M. Pollitt. 1997. The restructuring and privatization of Britain's CEGB – was it worth it? *Journal of Industrial Economics* 45: 269–303.

Schweppe, F., M. Caramanis, R. Tabors, and R. Bohn. 1988. *Spot pricing of electricity*. Norwell: Kluwer Academic Publishers.

Waddams, C. 2004. Spoilt for choice? The costs and benefits of opening UK residential energy markets. Working paper 04–1. Norwich: Centre for Competition and Regulation, University of East Anglia.

Wilson, R. 2002. Architecture of power markets. *Econometrica* 70: 1299–1340.

Wolak, F. 2000. An empirical analysis of the impact of hedge contracts on bidding behavior in a competitive electricity market. *International Economic Journal* 14: 1–39.

Wolfram, C. 1999. Measuring duopoly power in the British electricity spot market. *American Economic Review* 89: 805–826.

Electronic Commerce

Avi Goldfarb

Abstract

Electronic commerce is the exchange, distribution, or marketing of goods or services over the Internet. This article first reviews electronic commerce adoption across US industries.

While the Internet is used in most industries, it has had a profound impact only on a small number. Businesses that rely heavily on electronic commerce can be divided into four groups: retail, media, business-to-business and other intermediaries. Each of these is discussed. The article concludes with a discussion of some features of electronic commerce that are of special interest to economists: lower economic frictions, lower communication costs, lower marginal costs and rich data.

Keywords

Advertising; Bundling; Communication costs; Computer industry; Economic frictions; Electronic commerce; Internet, economics of the; Media; Menu costs; Price dispersion; Retail; Search costs; Switching costs

JEL Classifications

L81; L86

In this article, electronic commerce is defined as the exchange, distribution, or marketing of goods or services over the Internet.

There is, unfortunately, no standard definition used in the academic literature or the popular press. A broader definition would include all business facilitated by telephones, fax machines, televisions, and other technologies that are 'electronic'. This broad definition, however, becomes so large that it encompasses a substantial fraction of all economic activity since the 1950s. A narrower definition would focus only on items sold over the World Wide Web, the browser-enabled portion of the Internet. This definition omits much of the important business-to-business segment of electronic commerce and the numerous advertising-supported websites.

The definition used here encompasses a variety of ways in which businesses have used the Internet. The Internet is a worldwide network of computers that connect to each other using the communication protocols defined by TCP/IP. Electronic commerce includes businesses that have used the Internet to reach other businesses and to reach consumers directly. It includes businesses that sell products directly to their customers and businesses that function as intermediaries. This definition also includes businesses that operate only online, the online business of those that operate online and offline, and businesses that use the Internet but not as their primary business function.

Adoption of Electronic Commerce by Industry

While most attention has focused on those few businesses where the Internet is a fundamental part of their strategy, electronic commerce is just one aspect of business processes for most businesses. As of 2000, nearly 90 per cent of large US establishments used the Internet (Forman et al. 2002). Nearly all industries and cities had adoption rates well over 70 per cent. For the vast majority of these establishments, the Internet was used to send and receive email, to help automate some basic processes like inventory management, and/or for web browsing. This basic level of use was particularly important to establishments in rural areas (Forman et al. 2005). Overall, the impact on most industries, from nursing homes to construction to furniture manufacturing to petrol stations, has been limited. The Internet is used in day-to-day business activities, but it is a small piece in a much larger puzzle. Even in retail, the US Census reported that Internet sales (totalling $26.3 billion) were just 2.7 per cent of total US retail sales in the second quarter of 2006 (U.S. Census Bureau 2006b).

Still, a small portion of businesses have used the Internet to enhance business processes at a deep level. While little research has examined why some industries adopted quickly and others did not, it is the businesses that adopted quickly that get the majority of the attention. The Internet has had a profound effect on publishing, securities trading, some wholesaling, and some retailing (for example, books and computers). In particular, businesses that rely heavily on electronic commerce can be divided into four (not necessarily mutually exclusive) groups: retail, media, business-to-business (B2B), and other intermediaries.

Retail

Electronic commerce represents the introduction of a new sales channel. While the size of the online channel is still small relative to the entire retail sector, electronic commerce has had a large effect on some retail markets. According to the U.S.

Census, Internet sales made up over ten per cent of 2004 retail sales in two broad categories if online-only stores are included: electronics and appliance stores (that is, NAICS 443) and sporting goods, hobby, book, and music stores (that is, NAICS 451) (U.S. Census Bureau 2006a). Much of the literature on electronic commerce has focused on these categories, as well as motor vehicles and travel.

A new channel has the potential to create channel conflict. There is considerable evidence that consumers compare prices and options across channels (Prince 2006; Ellison and Ellison 2006). Forman et al. (2006) show that use of the online channel depends on local offline retail options. Also, Hendershott and Jie Zhang (2006) argue that manufacturers may face resistance from their retailers to setting up a direct online channel. They show that the benefits of selling directly to consumers (rather than though a retailer) depend on the relative online–offline search costs. The benefits of the online channel are largest for goods that are not widely available in retail stores (that is, high offline search costs) and for goods that do not need to be touched to assess quality (that is, low online search costs).

Media Websites

In addition to a new retail channel, the Internet has provided a new media outlet. This outlet has developed a market structure similar to the magazine industry (Goldfarb 2004). Media websites provide information to visitors and earn money (mostly) through advertising. In particular, entry is easy but distribution is difficult to achieve; concentration is largely determined by market size and distribution costs; large media conglomerates coexist with small niche players; and there is a high mortality rate. Online media appear to be particularly important to overcome local isolation (Sinai and Waldfogel 2004). The two-sided nature of the media market and the digital nature of the product mean that competition between media websites is different in nature from competition between online retailers.

Intermediaries

According to Alexa.com, six of the top seven most popular websites in October 2006 had roles as intermediaries: Yahoo, MSN, Google, MySpace, YouTube, and eBay. While these intermediaries may share features of media websites (Google) or retailers (eBay), their primary business is to facilitate online interactions. Without physical storefronts or displays, intermediaries help individuals (and firms) find each other online. Intermediaries allow people with heterogeneous tastes to find better matches in terms of media, products, and people (Scott Morton 2006).

Business to Business

Business-to-business (B2B) electronic commerce is a relatively under-researched area, perhaps because of the difficulties in obtaining data. Still, B2B transactions are many times the size of business-to-consumer transactions. Lucking-Reiley and Spulber (2001) summarize many of the key questions and opportunities in B2B electronic commerce including B2B exchanges, automatic ordering, and outsourcing. Some aspects of the Internet, such as asynchronous communication, may be particularly important for international B2B interactions. Many B2B applications can also be done on electronic data interchange (EDI) rather than the Internet.

Key Features of Electronic Commerce for General Economic Research

In addition to its widespread usage across industries and its profound impact on a small set of them, electronic commerce has a number of features that make it a particularly interesting area of study for economists.

Fewer Economic Frictions

The Internet reduces a number of economic frictions that are often cited as key contributors to

observed imperfections in markets. To the consumer, search and switching costs are reduced substantially. To the firm, menu and distribution costs may fall.

For consumers, the Internet makes it relatively easy to search through several retail options. Instead of having to walk from store to store, consumers can simply click from one company to another without leaving their desks. Furthermore, a number of intermediaries exist that reduce search costs even further. These 'shopbots' allows consumers to compare prices and features from several websites during a single keyword search. In addition to lower search costs, switching costs are also lower online than offline. It is not difficult to switch from one competitor to another. Much of the earliest research examining electronic commerce focused on why price dispersion persisted in this environment. Broadly speaking, this literature concluded that, all else equal, search and switching costs are lower online; however, firms created search and switching costs to overcome this challenge (Ellison and Ellison 2004). Consequently, there is still substantial price dispersion online. Still, low search costs do not mean zero search costs. Visibility matters to the long-term prospects of any business-to-consumer company. Many early Internet companies struggled because they misinterpreted low search costs as zero search costs, mistakenly assuming customers would arrive once they set up the website.

Firms also benefit from fewer frictions online. In particular, the menu costs of changing prices and updating product offerings are much lower online than offline. In addition to the reduction in menu costs, some firms benefit from lower distribution costs: for digital goods (namely, music, news and images) online distribution costs are near zero. Low menu costs combined with the digital nature of many online products allow for mass customization of products (Murthi and Sarkar 2003) and creative bundling, licensing, versioning and pricing strategies. Shapiro and Varian (1999) and Bakos and Brynjolfsson (1999) provide examples of a number of situations in which online firms are better able to match customers needs and therefore are better able to price discriminate.

Lower Communication Costs

The Internet reduces communication costs considerably. It provides an additional means of communication that creates new potential to interact with customers, suppliers and with other branches of the same firm. Internet communication differs from telephone communication in two primary ways. First, the marginal cost of communication is effectively zero, even over long distances. While establishing a connection is costly, each additional e-mail, web page viewed, and instant messaging interaction has no monetary cost to the communicator. Second, Internet communication is often asynchronous. Unlike telephone communications, the people communicating do not necessarily have to be available at the same time. This has many important applications. For example, it facilitates communication across time zones. Together, these features of Internet communication mean that geography may be less important online. Given access, people can communicate with any other person who has access, irrespective of location. Still, despite the substantial fall in long-distance communications costs, most online communication is local because social networks are local (Wellman 2001).

Lower Marginal Costs

Many goods sold over the Internet are digital in nature (for example, newspaper content, music, information). The marginal cost of replication for digital goods is near zero. Depending on the particular good, fixed costs may be high (software) or low (blogs). Shapiro and Varian (1999) discuss in detail the economics of goods with high fixed and low marginal costs. If fixed costs are high enough, this cost structure allows monopolists with broad flexibility in pricing, versioning and bundling policies. It also leads to substantial economies of scale and incentives to sell a broad scope of products. In markets with more than one player, this cost structure can lead to fierce competition and little profit. If fixed costs are low and entry is easy then prices should approach zero.

One misunderstood aspect of electronic commerce is that many Internet business models have not benefited from low marginal costs, and therefore have no cost advantage over offline competition. Low marginal costs apply only to digital goods

and services. In the late 1990s, many companies failed because their business models shipped heavy items to consumers. For example, taking orders for pet food and shipping it to customers involves very high marginal costs per item sold.

Rich Data

By definition, all online activity is digital. This means that it is relatively easy to record and store information on the behaviour of consumers and firms online. In contrast, it is extremely expensive to track all a shopper's activity in a typical offline store. Online, however, every item browsed and the time spent looking is easily recorded. This presents an opportunity for both firms and researchers. Firms can use this data to better understand their customers, which leads to more effective customization. Researchers can use this data to answer many questions that previously could not be answered due to data constraints. Online data has greatly enhanced of our understanding of a number of economic concepts including auctions (for example, Bajari and Hortacsu 2003), the economics of information (for example, Jin and Kato 2005), and social interactions (for example, Mayzlin and Chevalier 2006).

In summary, this article has identified some important features of electronic commerce and the some of the main areas of related economic research. Useful surveys of electronic commerce and related subjects include Scott Morton (2006), Hendershott (2007), and Ellison and Ellison (2005).

See Also

- ▶ Computer Industry
- ▶ Information Technology and the World Economy
- ▶ Internet, Economics of the
- ▶ Price Dispersion

Bibliography

Bajari, P., and A. Hortacsu. 2003. The winner's curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions. *RAND Journal of Economics* 34: 329–355.

Bakos, Y., and E. Brynjolfsson. 1999. Bundling information goods: Price, profits, and efficiency. *Management Science* 45: 1613–1630.

Ellison, G., and S.F. Ellison. 2004. *Search, obfuscation, and price elasticities on the Internet*. Working paper, no. 10570. Cambridge, MA: NBER.

Ellison, G., and S.F. Ellison. 2005. Lessons about markets from the Internet. *Journal of Economic Perspectives* 19(2): 139–158.

Ellison, G., and S.F. Ellison. 2006. *Internet retail demand: Taxes, geography, and online-offline competition*. Working paper no. 12242. Cambridge, MA: NBER.

Forman, C., Ghose, A., and Goldfarb, A. 2006. *Geography and electronic commerce: measuring convenience, selection, and price*. Working paper no. 06–15. New York: NET Institute.

Forman, C., Goldfarb, A., and Greenstein, S. 2002. *Digital dispersion: An industrial and geographic census of commercial Internet use*. Working paper no. 9287. Cambridge, MA: NBER.

Forman, C., A. Goldfarb, and S. Greenstein. 2005. How did location affect adoption of the commercial Internet? Global village vs. urban leadership. *Journal of Urban Economics* 58: 389–420.

Goldfarb, A. 2004. Concentration in advertising-supported online markets: An empirical approach. *Economics of Innovation and New Technology* 13: 581–594.

Hendershott, T., eds. 2007. *Handbook of economics and information systems*. Amsterdam: North-Holland.

Hendershott, T., and Jie Zhang. 2006. A model of direct and intermediated sales. *Journal of Economics & Management Strategy* 15: 279–316.

Jin, G.Z. and Kato, A. 2005. Price, quality, and reputation: Evidence from an online field experiment. *RAND Journal of Economics* 37:983–1004. (forthcoming).

Lucking-Reiley, D., and D.F. Spulber. 2001. Business-to-business electronic commerce. *Journal of Economic Perspectives* 15(1): 55–68.

Mayzlin, D., and J.A. Chevalier. 2006. The effect of word-of-mouth on sales: Online book reviews. *Journal of Marketing Research* 43: 345–354.

Murthi, B.P.S., and S. Sarkar. 2003. The role of the management sciences in research on personalization. *Management Science* 49: 1344–1362.

Prince, J. 2006. The beginning of online/retail competition and its origins: An application to personal computers. In the International Journal of Industrial Organization25 139–156 (forthcoming).

Scott Morton, F. 2006. Consumer benefit from use of the Internet. In *Innovation policy and the economy*, ed. A.B.. Jaffe, L. Lerner, and S. Stern, vol. 6. Cambridge, MA: MIT Press.

Shapiro, C., and H.R. Varian. 1999. *Information rules: A strategic guide to the network economy*. Boston: Harvard Business School Press.

Sinai, T., and J. Waldfogel. 2004. Geography and the Internet: Is the Internet a substitute or a complement for cities? *Journal of Urban Economics* 56: 1–24.

U.S. Census Bureau. 2006a. E-Stats, 25 May. Online. Available at: http://www.census.gov/eos/www/papers/2004/2004reportfinal.pdf. Accessed 13 Jan 2007.
U.S. Census Bureau. 2006b. Quarterly retail e-commerce sales, 2nd quarter 2006.
U.S. Census Bureau News. Online. Available at: http://www.census.gov/mrts/www/data/html/06Q2.html. Accessed 13 Jan 2007.
Wellman, B. 2001. Computer networks as social networks. *Science* 29: 2031–2034.

Elites and Economic Outcomes

Elise S. Brezis and Peter Temin

Abstract

Elites are a necessary part of economic activity. It therefore matters how elites are recruited and how they act. History is full of examples of elites that have acted well and also badly. Modern research has examined the training of elites, recruitment schemes and incentives for elites to discover how they can be used to promote, rather than impede, economic growth. The literature has also emphasized the effect of elite interconnection and elite recruitment on social mobility; it has shown that the standardization of elite education over the years may lead to uniformity and the creation of a transnational oligarchy.

Keywords

Circulation of elites; Class; Corruption; Cultural capital; Economic growth; Education; Elites and economic outcomes; Globalization; Human capital; Inequality; Iron law of oligarchy; Meritocracy; Power elite; Social capital; Social mobility; Symbolic capital; Technical change

JEL Classifications

N8; D3; Z13

A ruling elite (from the Latin *eligere*, 'to elect') is a small, dominant group that enjoys the power of decision in the various sectors of the economic and social organization of a state. It includes the bureaucrats and civil servants who rule the macro-environment; the political elite that governs and operates the executive, legislative and judicial structures; and the business elite. Non-ruling elites include the members of the media, academia and the intelligentsia.

Even in a democratic regime in which the power is meant to reside in the *demos* ('the people'), power is really concentrated in the hands of a few. All political organizations, even democracies, tend towards domination by an oligarchy, which Mills (1956) called the *power elite*. This is the *iron law of oligarchy* as stated by Michels (1915). This stratification of society based on the accumulation of decisionmaking power therefore differs from the familiar stratification based on income and economic means, or on ownership of the factors of production as emphasized by Marx.

The effects of elite actions on the economy operate through several channels: economic growth and development; social mobility; inequality; and the political system, which in turn affects the economy. The characteristics that affect these economic realms are (*a*) the extent of the intertwining and inter-connections of elites; and (*b*) the stability and recruitment of the elite.

Elites' Interconnections

The ruling elite can display unity and collusion, acting as a monolithic group, or it can be fragmented and characterized by dissociation and diversification of power, a 'polyarchy' that permits competition among its members.

The elite in non-democratic polities displays unity, has unlimited political and economic power, and typically acts on behalf of its own interests. But democracy should a priori impose some control on the power of the ruling elite. Indeed, Schumpeter (1954) claimed that the democratic process permits 'free competition among would-be leaders for the vote of the electorate' and that the masses can choose between various elites. In contrast, classical elite theorists such as Mosca (1939); Pareto (1935), Michels and Mills emphasized that there can be collusion even in

democracies. Numerous elites may not be mutually competitive and may not control and balance each other; instead, they may be intertwined as a unanimous, cohesive power elite.

Economic Consequences of the Extent of Interconnection

Inequality

The elite's plurality and competition ensures its responsiveness to the demands of the public, while a consensual elite might use its power for its own interests. Etzioni-Halevy (1997) claims that a unified elite does not use its power to reduce inequality and promote the development of a more egalitarian society, due to common recruitment and common interests. It is the plurality and differentiation of the members of the elite that enables them to countervail each others' power and to increase their responsiveness to the will of public. In consequence, elite homogeneity might actually increase the gap between the elite and the masses.

When the political elite controls wealth and the main factors of production, then elite and class stratifications coincide, and consequently power and wealth are in the hands of the same happy few. Engerman and Sokoloff (1997) showed that members of the elite who have power and wealth establish institutions that serve their own interests and exclude the masses from benefits. In consequence, inequality persists through institutional development in the elite's own favour. Justman and Gradstein (1999) added that elite unity leads to greater inequality through regressive redistribution policy. A power elite that controls wealth may refrain from investment in human capital of the majority because education would increase the latter's political voice and weaken the elite's hold on power (Easterly 2001); yet in some cases, the elite deliberately decides to forfeit power by investing in human capital as a consequence of a cost–benefit analysis (Bourguignon and Verdier 2000).

The extent of elite unity can be endogenously determined (Sokoloff and Engerman 2000), and elite unity can also be affected by revolutions, wars and economic growth. Justman and Gradstein (1999) argue that economic growth dilutes the power of the elite by broadening political participation and reducing inequality.

Economic Growth

A strong interconnection among elites has the consequence that all sectors of the economy are ruled by a group that thinks in a monolithic way. Two lines of thoughts have related a monolithic group to economic growth. The first one underlines that a monolithic group leads to the stagnation of ideas and attitudes, which in turn may prevent the adoption of major technological breakthroughs (Bourdieu 1977). The lack of competition in a monolithic powerful group also generates corruption, with harmful consequences for growth.

The second line of thought argues that wealthy elites with enough political power to block changes will not accept adopting institutions that would enhance growth, since they might hurt them. Acemoglu et al. (2001) developed this line of thought in relation to colonial impacts, showing that, wherever colonial governments were composed of few elite members, economic progress was reduced.

Following the same line of reasoning, Acemoglu and Robinson (2000) and Gradstein (2007) stressed that elite plurality, in which the political and economic elites are separate, explains the adoption of political franchise and industrialization in western Europe; while 19th-century eastern Europe, where elite unity was strong, did not adopt growth-enhancing institutions, since its elites held on to their wealth and power.

Paradoxically, in countries in which the elite was united and consensual, with common aims, the transition to capitalist production in the 1990s took place without violence, as in Poland and the Czech Republic. In contrast, wherever the elite was divided and fragmented, there were conflicts, especially on the ethno-nationalist level, as in Yugoslavia and Romania (Pakulski 1999).

Recruitment and Training of Elites

Plato claimed that government should be in the hands of the most able members of society, that is,

the *aristocracy* (Greek for 'rule by the best'), a term that became pejorative and was later changed to *meritocracy* (coined by Young 1958). Pareto argued that a stable economic system needs a *circulation of elites*, so that the most capable and talented are in the governing class. He stressed that the quality of the ruling class can be maintained only if social mobility is allowed, so that the non-elite has the possibility of entering the elite: 'History is a cemetery of aristocracies' (Pareto 1935). His theory may be viewed as a sort of social Darwinism in which mobility is needed, just as evolution relied upon competition and selection.

For millennia, recruitment of the Western elite was based on social inheritance and was carried out via heredity, nepotism and violence. Hereditary monarchy was considered the most legitimate means of recruitment for rulers, and the upper elite was made up of wealthy large landowners, an *état de fait* considered normal in agrarian societies. Nevertheless, there were some channels of entrance into the elite, such as military prowess and exploits or involvement in government finance (Brezis and Crouzet 2004).

In democracies, the political elite came to be recruited mainly by election. Yet for a long time, the franchise was not for all. Big landowners and members of the upper middle class were the overwhelming majority in parliaments and cabinets, even though some prominent business people entered the political elite. Only in the late 19th century did members of the lower middle class and working class enter the political elite.

From the 19th century onwards, the circulation of the business elite took two differing yet concurrent paths. The first was that economic growth led to spurts of new firms and the decline of others, allowing a new business elite to emerge (Schumpeter 1961). The second path was the rise of the professions, with competitive and meritocratic exams that led to circulation of elites (Perkin 1978). After the Second World War, the elite was mainly recruited through education into elite universities to which admission started to be conferred following success at meritocratic exams.

Economic Consequences of the Recruitment of Elites

Social Mobility in the Economy

Prior to recruitment through meritocracy, social mobility, and in particular the potential for non-elite members to enter the elite, was low. Temin (1999a, b) showed that today, as in the 1900s, and despite meritocracy, the American economic elite is composed almost entirely of white Protestant males who have been educated for the most part in Ivy League colleges. Although in 1900 the political elite was quite similar to the business elite, today the former is more diversified; the political elite has changed in its recruitment, while the economic elite has not. In other words, minorities have not penetrated the economic elite in the United States (see also Friedman and Tedlow 2003, which summarizes studies on US elite mobility, and Foreman-Peck and Smith 2004 on British elites).

Recruitment to a university through meritocratic entrance exams, does not, indeed, lead to enrolment from all classes of society according to distribution or ability, nor does it necessarily lead to the admission of the most talented. Recruitment by entrance exam still encompasses a bias in favour of elite candidates because this type of exam requires a pattern of aptitude and thinking that favours candidates from an elite background. All elite positions may be open to all applicants with the right qualifications, but they are more accessible to those with specific social, cultural and symbolic capital (Arrow et al. 2000). Thus the power elite maintains its status and power by a *strategy of distinction*, or a cultural bias that is necessary for accessing it (Bourdieu 1977). A small difference in culture and education leads to narrow recruitment, and in turn to class-based stratification in the recruitment of the elite, despite meritocratic selection for universities (Brezis and Crouzet 2006).

The relationship between mobility and the political system, as emphasized by Pareto, has been analysed by sociologists. For instance, Lengyel (1999) showed that circulation in the elite occurs at times of political upheavals and revolutions: the existing elite is eliminated and

replaced by a new one. The first-generation members of the elite following a political change have neither specific training and education nor specific origin; they are the trailblazers, the entrepreneurs who seized power on the strength of their competence. In the next generation, the elite becomes narrowly recruited from the best educated, and members are selected mostly by training and education. The elite returns to an occupational specialization, similar to the meritocratic profession criterion of earlier industrialization (Perkin 1978).

Economic Growth

A crucial element of economic growth is that the recruited elite be of the highest quality. Countries in which elites are recruited in a non-meritocratic way face the problem of the quality of their elites. However, the prevalence of meritocratic recruitment does not necessarily lead to the selection of the best ruling elites. Brezis and Crouzet (2006) argue that, when a country faces only mild technological and structural changes, the narrow recruitment, due to meritocracy, optimally fulfils its purpose, since the cultural bias of the elites is an advantage in the given type of technology. However, at times of major changes in technology, elites recruited this way are not the best for adopting new technologies.

Moreover, the homogeneity of the recruitment of elites through similar curricula leads to convergence of views; this, in turn, leads to a monolithic elite, which, as we have claimed above, may have negative consequences for economic growth.

Conclusion

In this short article, we have summarized the modern research that has examined recruitment schemes and incentives for elites to discover how they can be used to promote, rather than impede, economic growth. There is also an entire economic history literature that has enriched us with a wealth of knowledge on the business elite. The main works in this literature are by Cassis (1997); Crouzet (1999) and Lachmann (2000).

The literature cited herein seems to show that the structure of this small group called the elite has numerous effects on the world economy. In the opposite direction, globalization will also affect the elite, as we are now facing a globalization of education of the elite.

In its first wave, globalization of education will probably create a new collection of elites and elicit some changes, yet the unity and uniformity of the elite will be even greater, not only at the national level but also at the global level. National elites will be replaced by a worldwide elite, along with uniformity in culture and education. We will face an international technocratic elite with its own norms, ethos, and identity, as well as its private clubs like the Davos World Economic Forum – a transnational oligarchy.

See Also

- ▶ Economic Growth, Empirical Regularities in
- ▶ Income Mobility
- ▶ Pareto, Vilfredo (1848–1923)
- ▶ Social Status, Economics and

Bibliography

Acemoglu, D., S. Johnson, and J.A. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–1401.

Acemoglu, D., and J. Robinson. 2000. Political losers as a barrier to economic development. *American Economic Review* 90: 126–144.

Arrow, K., S. Bowles, and S. Durlauf. 2000. *Meritocracy and economic inequality*. Princeton: Princeton University Press.

Bourdieu, P. 1977. *Reproduction in education, society and culture*. London: Sage.

Bourguignon, F., and T. Verdier. 2000. Oligarchy, democracy, inequality and growth. *Journal of Development Economics* 62: 285–313.

Brezis, E.S., and F. Crouzet. 2004. Changes in the training of the power elites in Western Europe. *Journal of European Economic History* 33: 33–58.

Brezis, E.S., and F. Crouzet. 2006. The role of higher education institutions: Recruitment of elites and economic growth. In *Institutions and economic growth*, ed. T. Eicher and C. Garcia-Penalosa. Cambridge, MA: MIT Press.

Brezis, E.S., and P. Temin, eds. 1999. *Elites, minorities, and economic growth*. Amsterdam: Elsevier.

Cassis, Y. 1997. *Big business. The european experience in the twentieth century*. Oxford: Oxford University Press.
Crouzet, F. 1999. Business dynasties in Britain and France. In *Elites, minorities, and economic growth*, ed. E.-S. Brezis and P. Temin. Amsterdam: Elsevier.
Easterly, W. 2001. The middle class consensus and economic development. *Journal of Economic Growth* 6: 317–335.
Engerman, S.L., and K.L. Sokoloff. 1997. Factor endowments, institutions, and differential paths of growth among New World economies: A view from economic historians of the United States. In *How latin America fell behind: Essays on the economic history of Brazil and Mexico, 1800–1914*, ed. S. Haber. Stanford: Stanford University Press.
Etzioni-Halevy, E. 1997. *Classes and elites in democracy and democratization*. New York: Garland.
Foreman-Peck, J., and J. Smith. 2004. Business and social mobility into the British elite 1870–1914. *Journal of European Economic History* 55: 485–518.
Friedman, W.A., and R.S. Tedlow. 2003. Statistical portraits of American business elites: A review essay. *Business History* 45: 89–113.
Gradstein, M. 2007. Inequality, *democracy* and the protection of property rights. *Economic Journal* 117: 252–269.
Higley, J., and M. Burton. 1989. The elite variable in democratic transitions and breakdowns. *American Sociological Review* 54: 17–32.
Justman, M., and M. Gradstein. 1999. Industrial Revolution, political transition and the subsequent decline in inequality in nineteenth-century Britain. *Explorations in Economic History* 36: 109–127.
Lachmann, R. 2000. *Capitalist in spite of themselves: Elite conflict and economic transition*. Oxford: Oxford University Press.
Lengyel, G. 1999. Two waves of professionalization of the Hungarian economic elite. In *Elites, minorities, and economic growth*, ed. E.S. Brezis and P. Temin. Amsterdam: Elsevier.
Michels, R. 1915. *Political parties*, 1959. New York: Dover.
Mills, C.W. 1956. *The power elite*. New York: Oxford University Press.
Mosca, G. 1939. *The ruling class*. New York: McGraw.
Pakulski, J. 1999. Elites, ethnic mobilization and democracy in postcommunist Europe. In *Elites, minorities, and economic growth*, ed. E.S. Brezis and P. Temin. Amsterdam: Elsevier.
Pareto, V. 1935. *The mind and society*. New York: Harcourt Brace.
Perkin, H. 1978. The recruitment of elites in British society since 1800. *Journal of Social History* 12: 222–234.
Sokoloff, K., and S. Engerman. 2000. Institutions, factor endowment and paths of development in the New World. *Journal of Economic Perspectives* 14(3): 217–232.
Schumpeter, J.A. 1954. *Capitalism, socialism and democracy*. London: Routledge.
Schumpeter, J.A. 1961. *Theory of economic development*. New York: Oxford University Press.
Temin, P. 1999a. The American business elite in historical perspective. In *Elites, minorities, and economic growth*, ed. E.S. Brezis and P. Temin. Amsterdam: Elsevier.
Temin, P. 1999b. The stability of the American business elite. *Industrial and Corporate Change* 8: 189–210.
Young, M. 1958. *The rise of meritocracy, 1870–2033*. London: Thames & Hudson.

Ellet, Charles, Jr. (1810–1862)

Robert B. Ekelund

Keywords

Differential calculus; Duopoly; Ellet, C., Jr; Firm, theory of; Market area analysis; Monopoly; Price discrimination

JEL Classifications

B31

American engineer and economic theorist, Ellet was born on 1 January 1810 at Penn's Manor, Pennsylvania, and died on 21 June 1862, a victim of the Civil War. Ellet grew up on a family farm but showed little inclination for agriculture: at age 17 he joined a surveying crew. With no formal education or training, he soon became an assistant engineer to Benjamin Wright, chief engineer of the Chesapeake and Ohio Canal. With ability and hard work Ellet taught himself mathematics and French, earning the respect of influential engineers. Letters of introduction to Lafayette and the American ambassador helped secure Ellet a place at the Ecole des Ponts et Chaussées, Dupuit's alma mater, in 1830. On his return to America in 1832 Ellet became the premier suspension bridge designer in America, building in 1849 the (then) longest suspension bridge in the world across the Ohio River at Wheeling. Colonel Ellet designed, constructed and commanded the ram fleet of the Union forces at the naval battle at Memphis, Tennessee. He

died as a result of a wound received in the heat of that battle.

Ellet spent most of his professional life as an engineer, but, in one major work and in a number of contributions to the *Journal of the Franklin Institute* between 1840 and 1844, he significantly advanced the economic theory of monopoly, input selection, spatial economics, benefit–cost theory and econometric estimation. All Ellet's contributions were facilitated by the use of the differential calculus, which permitted him to express the simple theory of the firm, and some of its extensions, in mathematical terms. In his *Essay on the Laws of Trade* (1839) Ellet established the demand curve for a monopoly railroad with distance as a variable. Utilizing firstorder conditions and solving for the gross toll on passenger traffic, Ellet demonstrated that the profit-maximizing toll would be equal to one-half the costs of transportation added to a constant quantity, a well-known result.

Ellet considered not one monopoly model but a multiplicity of them, including those dealing with freight transport, duopoly conditions and the principles of monopoly price discrimination. Further, Ellet's particular insights into simple and discriminatory pricing systems led him to provide, with distance as a variable, an amazingly complete mathematical and graphical analysis of the impact of changes in the pricing system upon the market area served by a profit-maximizing railroad (1840a). In this important contribution to market area analysis Ellet argued that a set of (constrained) discriminatory tolls inverse to distance, in contrast to tolls proportional to distance, could be devised whereby all interested parties (management, shippers, the state) could be made better off. In a series of papers (1842–4) Ellet extended his theoretical analysis of inputs and input selection (1839) to one of the earliest attempts to develop, empirically specify and test a theoretical cost function. Utilizing a 'law' of costs which included his selected determinants of annual total railway costs, Ellet estimated the empirical dimensions from data collected from the mid-1830s. He then reaffirmed the power of his initial equation with new and supplementary data.

In all, the calibre and completeness of Ellet's theoretical and empirical inventions would not compare unfavourably with those of von Thünen, Cournot, Dupuit or Lardner. Ellet, who was primarily an engineer, was America's best representative among the pioneer contributors to scientifically oriented economics in the 19th century.

Selected Works

1839. *An essay on the laws of trade in reference to the works of internal improvement in the United States.* Richmond. Reprint from the 1st ed., New York: Augustus Kelley, 1966.

1840a. The laws of trade applied to the determination of the most advantageous fare for passengers on railroads. *Journal of the Franklin Institute* 30: 369–379.

1840b. A popular exposition of the incorrectness of the tariffs of tolls in use on the public improvements of the United States. *Journal of the Franklin Institute* 29: 225–232.

1842–4. Cost of transportation on railways. *Journal of the Franklin Institute,* various issues.

Bibliography

Baumol, W., and S.M. Goldfeld (eds.). 1968. *Precursors in mathematical economics: An anthology.* London: London School of Economics and Political Science.

Calsoyas, C.D. 1950. The mathematical theory of monopoly in 1839: Charles Ellet, Jr. *Journal of Political Economy* 58: 162–170.

Ekelund Jr., R.B., and D. Hooks. 1972. Joint demand, discriminating two-part tariffs and location theory: An early American contribution. *Western Economic Journal* 10: 84–94.

Viner, J. 1958. *The long view and the short: Studies in economic theory and policy.* New York: Glencoe.

Ely, Richard Theodore (1854–1943)

A. W. Coats

Keywords

American Economic Association; Commons, J.R.; Ely, R.T.; German Historical School; Institutionalism; Knies, K.G.A.

JEL Classifications
B31

Ely was born in Ripley, New York, on 13 April 1854 and died at Old Lyme, Connecticut, on 4 October 1943.

Ely's long and vigorous career epitomizes the general proposition that an economist can exert a major constructive influence on his subject and profession even though his original contribution to economic theory is negligible. A highly effective teacher and maker of careers for his former students; prolific author of popular articles, scholarly volumes, and publications series; organizer and fund-raiser for major research projects; founder of various academic institutes and associations; leader or participant in numerous reform societies; and centre of innumerable controversies, Ely was the most widely known, even notorious, economist in the USA around the turn of the 20th century.

After a brief spell as a country schoolteacher and a preliminary year at Dartmouth College, Ely graduated from Columbia College in 1876 and was awarded a three-year fellowship to study philosophy in Germany. He soon switched to political economy, came under the influence of Karl Knies at Heidelberg, where he obtained a Ph.D., summa cum laude, in 1878, and later attended Adolph Wagner's lectures in Berlin. Returning to the USA he was unemployed for more than a year before his appointment, initially on a half-time basis, at Johns Hopkins, where he taught from 1881 to 1892. He then moved to Wisconsin, founding an outstanding school of Economics, Political Science and History including such luminaries as F.J. Turner, E.A. Ross, and J.R. Commons. A unique collaboration developed between the social scientists and the state legislators, especially under the La Follette governorship, which pioneered major social and economic reform legislation. In 1925 Ely took his Institute for Research in Land Economics and Public Utilities, founded in 1920, from Madison to Northwestern University, and remained there until 1932, when he launched a new, but impoverished Institute for Economic Research in New York City. Eventually hit by the depression, Ely was forced to depend on the support of friends and former students as he completed his autobiography and failed to complete a massive history of American economic thought initiated 50 years earlier.

An ardent Christian Socialist and outspoken critic of laissez-faire individualism and 'old school' English classical economics, Ely delighted social reformers and outraged conservatives by his writings on such controversial current topics as socialism and the American labour movement. Prone to emotional overstatement and careless in exposition, his public pronouncements and reputation frequently embarrassed the aspiring young professional economists with whom he founded the American Economic Association, in 1885, and for a time discouraged some moderate and conservative economists from joining. Although Ely's original draft prospectus had been rejected, and the association's original constitution was toned down, and then dropped, the organization hovered uneasily between missionary evangelism and scholarly objectivity until he was obliged to relinquish his secretaryship in 1892.

Two years later, at Wisconsin, Ely's fellow professionals rallied around him when he was denounced for preaching socialism and encouraging strikes, and, although he was completely exonerated in a 'trial' that attracted national attention, Ely gradually became more conservative. Ironically, in the 1920s his institute was attacked, no doubt unfairly, as a tool of the public utilities, and was referred to disparagingly in a report on professional ethics by a committee of the American Association of University Professors, in 1930.

During his long lifetime Ely wrote extensively on an extraordinarily wide variety of topics, often in a popular and journalistic fashion. Nevertheless, he repeatedly opened up new research topics that were developed by his colleagues and former students – for example, in labour history, state taxation, land economics, and natural resources – and his various textbooks, especially the multi-edition *Outlines of Economics* which sold 350,000 copies, were both widely used and highly regarded.

At Wisconsin he helped to launch the American Association for Labor Legislation, of which he became President, and raised private resources to finance John R. Commons's massive *Documentary History of American Society* (11 vols, 1910–11). He served as President of the American Economic Association in 1900–1901.

Ely was a stimulating teacher whose ideas formed a direct link between the doctrines of the German Historical School and American institutionalism, a link most clearly evident in his neglected two-volume study of *Property and Contract in their Relations to the Distribution of Wealth* (1914). Many of his students went on to distinguished careers in academic and/or public life. He was undoubtedly an outstanding academic entrepreneur, and his contribution to the American Economic Association is recognized in its annual invited Richard T. Ely lecture, which was inaugurated in 1963.

Selected Works

1883. *French and German socialism in modern times.* New York: Harper & Brothers. Reprinted, 1911.
1884a. *The past and the present of political economy.* Baltimore: N. Murray for Johns Hopkins.
1884b. *Recent American socialism.* Baltimore: N. Murray for Johns Hopkins. Reprinted, 1885.
1886. *The labour movement in America.* New York: T.Y. Crowell Co. New edn, revised and enlarged, New York: Macmillan Co, 1905.
1888a. *Taxation in American states and cities*. New York: T.Y. Crowell Co.
1888b. *Social aspects of christianity.* Boston: W.L. Greene and Co.
1889a. *An introduction to political economy.* New York: Chautauqua Press. New and revised edn, New York: Eaton and Mains; Cincinnati: Jennings and Pye, 1901.
1889b. *Social aspects of christianity and other essays.* New York: T.Y. Crowell Co. Reprinted, 1895.
1893. *Outlines of economics.* Meadville, Pennsylvania and New York: Flood and Vincent. 6th edn with Ralph Hess, New York: Macmillan Co., 1938.
1894. *Socialism: An examination of its nature, its strength, its weakness. With suggestions for social reform*. London: S. Sonnenschein Co.; Newyork and Boston: T.Y. Crowell Co.
1900. *Monopolies and trusts.* New York: Macmillan Co. Reprinted, 1912.
1903. *Studies in the evolution of industrial society.* New York/London: Macmillan Co. Reprinted, 1918.
1914. *Property and contract in their relations to the distribution of wealth.* 2 vols. New York: Macmillan Co. Reprinted, 1922.
1924. (With E.W. Morehouse.) *Elements of land economics.* New York: Macmillan Co. Reprinted, 1932.
1928. (With G.S. Wehrwein.) *Land economics*. Ann Arbor: Edwards Bros. Revised edn, Madison: University of Wisconsin Press, 1964.
1938. *Ground under our feet: An autobiography.* New York: Macmillan Co.

Bibliography

Rader, B.G. 1966. *The academic mind and reform: The influence of Richard T. Ely in American life*. Lexington: University of Kentucky Press.

Emergence

Yannis M. Ioannides

Abstract

With its philosophical pedigree and its use especially among life scientists and science writers since the early 1990s, the term 'emergence' in economics is more evocative than precise, reflects influence from physics and biology, and is now associated with phenomena where economic structures evolve into qualitatively different forms. These exhibit properties that are *emergent* in that they apply at an aggregate level but lack individual

analogues and therefore are not describable at the individual level. This article emphasizes applications that possess firm economic foundations, from the evolution of patterns in international trade to the establishment of a common currency.

Keywords

Aggregation; Autarky; Class; Cobb–Douglas functions; Complementarities; Congestion; Division of labour; Economic geography; Elasticity of substitution; Emergence; Fiat money; Financial market globalization; First fundamental theorem of welfare economics; Foreign aid; Herding; Innovation; International currencies; International currency; International finance; International trade; Invisible hand; Mill, J. S.; Multiple equilibria; Neighbours and neighbourhoods; Poverty traps; Power laws; Residential segregation; Specialization; Spontaneous order; Stochastic stability theory; Tipping points; Trade costs; Urban agglomeration; Urbanization

JEL Classifications

D85

Having acquired widespread use among life scientists and science writers since the early 1990s, the term 'emergence' in economics is more evocative than precise, reflects influence from physics and biology, and has come to be associated with phenomena involving evolution of economic structures into qualitatively different forms. These phenomena exhibit properties that are emergent in the sense that they are novel and apply at an aggregate more 'complex' level but lack individual analogues and therefore are not describable at, or reducible to, the individual level. A good case in point is the statement that consciousness is an emergent property of the brain. The notion of emergence originates in the philosophy of science, with John Stuart Mill being an important precursor (see Stanford Encyclopedia of Philosophy 2002).

This article reviews, albeit selectively, the recent usage of the term by emphasizing applications with predominantly economic phenomena where emergence of macroscopic properties may be elucidated by means of economic arguments. These range from neighbourhood tipping and evolution of patterns in international trade to emergence of urban structure and the establishment of norms and institutions and of a common currency, among many others.

More generally, emergent properties or behaviours have been studied in a variety of circumstances in nature, such as emergence of differentiated behaviour in colonies of animals, of herding behaviour in organizations and markets, of specialization of individuals into occupations and of cities and of regions and countries in specific products, of groups of biological cells in multicellular biological organisms and even of groups of processors in computer simulations involving cellular automata (see Holland 1998). The World Wide Web is an example of a decentralized engineering system that is continuously being modified by human initiatives in the form of actions by individuals and firms. The web has not been deliberately designed and no central organization administers how different sites are linked to others. Some of the properties of the graph topology of the web may be termed as *emergent*, such as that the number of links pointing to each page follows approximately a power law, with a few pages being pointed to by many others and most others seldom, and the fact that any pair of pages can be connected to each other through a relatively short chain of links in the average.

The presence of 'emergence' within the vocabulary of economists does suggest some interplay with multidisciplinary research by scientists who have been associated with the Santa Fe Institute (http://www.santafe.edu). To quote from Kauffman (1995, p. 24), an alternative definition of emergence is that '[t]he whole is greater than the sum of its parts'. And 'life itself is an emergent phenomenon . . . arising as the molecular diversity of a prebiotic chemical system increases beyond a threshold of complexity. If true, then life is not located in the property of any single molecule – in the details – but is a collective property of systems of interacting molecules.' The entirety of complex

molecules together is able to reproduce and evolve, a 'stunning property'.

Blume and Durlauf (2001) argue that emergence plays an important role even within the body of neoclassical economics proper. For example, the extent to which macroeconomics is a distinct discipline from microeconomics would be explained by emergent properties as alluded to by the statement 'aggregation is not summation' (see Kirman 1992). Consider, within microeconomics and general equilibrium theory, the metaphor of the invisible hand of the market (which goes back to Adam Smith), whereby individuals' pursuit of their own selfish aims leads to social outcomes that obey important social properties. Under certain conditions, after markets have brought about an equilibrium, it is impossible to make anyone better off without making someone worse off. Thus, the first fundamental theorem of welfare economics is an emergent property of social outcomes. However, the more modern work on emergence in economics has emphasized emergence of patterns. Similarly, Hayek's concept of spontaneous order may be considered an instance of emergence.

There are numerous other contexts where emergence has been alleged to occur. This article explores a number of examples of emergence that are limited to social and economic settings. They underscore the scope of the concept of emergence in such settings. As discussed earlier, there are many other contexts in socioeconomic settings and beyond, ranging from computation to the life sciences.

Emergent Social Interconnections

Suppose that a society consists of I individuals, where I is large, where any two individuals may be linked in a way that allows for communication, social relations, or social interactions. Let pk denote the probability that each individual is connected with exactly k other individuals. A literature going back to Erdös and Renyi (1960) and continuing at the time of writing up to Newman et al. (2001) has studied the topological properties of the (random) graph formed by the agents as nodes and connections between agents as edges when each agent's connections with other follows a given distribution p_k and the number of agents is large. According to Newman et al. (2001), depending upon whether the quantity $E[k^2] - 2E[k]$ is greater than or equal to 0, or falls below 0, there emerges, as I tends to infinity, a *proportion* of all individuals being interconnected, or, alternatively, the economy consists of different groups of finite sizes. In other words, the social structure undergoes a *phase transition* when this quantity exceeds 0: a giant interconnected component emerges. Intuitively, starting from a connected component of the graph, consider adding a new edge that connects with a previously isolated node of degree k. Doing so will change the number of nodes on the boundary of the connected component by $-1 + (k - 1) = k - 2$. The likelihood that a node is on the boundary of the connected component is proportional to k. The expected change in the number of nodes on the boundary when an additional node is connected is given by $\sum_i k_i (k_i - 2)/\sum_i k_i$. If this quantity is negative, then the number of nodes on the boundary decreases and therefore the connected component will stop growing. If it is positive, on the other hand, then the number of boundary nodes will grow and the connected component will grow, limited only by the size of the network.

In the simple case of the Erdös and Renyi random graph, where the number of connections is proportional to the number of individuals, the phase transition occurs when the factor of proportionality is equal to 1/2 and the corresponding average number of connections per person is equal to 1. Below this value, there are too few edges and the components of the random graph are small; above that value, a proportion of the entire graph belongs to a single, *giant* component. In this case, emergence of a qualitatively different social structure depends on the value of a single parameter (Kirman 1983; Ioannides 1990; Durlauf 1997). Individual behaviour that leads to a law for the number of individuals' connections does not necessarily imply the same macroscopic outcome in all circumstances. Similarly, social outcomes are not described by means of mere summation of individual actions; aggregation is

not summation (Kirman 1992). Kauffman (1995, p. 57) invokes this in the context of autocatalytic reactions and goes as far as seeing this 'as a toy version of phase transition that I believe led to the origin of life'.

Patterns of Residential Segregation

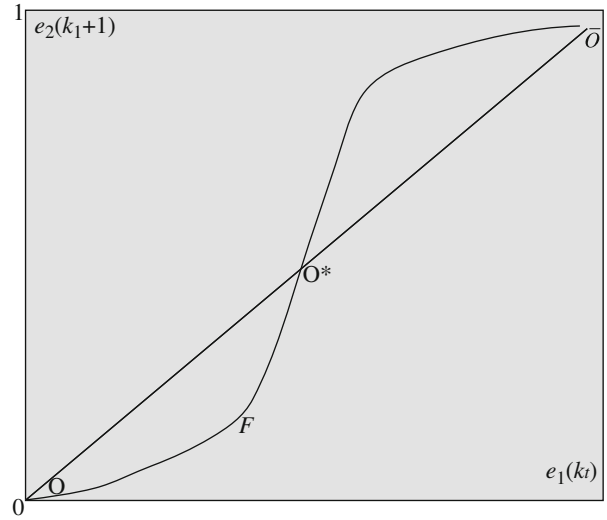
Now we turn to a description of neighbourhood tipping, which is originally due to Thomas C. Schelling (1978) and has been adapted here from recent works. Suppose that individual i is white and would live in a neighbourhood provided that the percentage of whites among her neighbours, $\omega \in [0, 1]$, is at leastw_i , $\omega \geq w_i$. She moves out otherwise. Individuals differ in terms of preference characteristic w_i, which is assumed to be distributed in a typical neighbourhood according to $F(\omega)$, when the analysis starts. For any neighbourhood with a share of white residents equal to u, the percentage of white individuals who would find living there acceptable are those with $w < \omega$. Their share is given by the value of the cumulative distribution function at ω, $F = F(\omega)$.

In Fig. 1, let the horizontal axis e_1 denote u and w_i, the vertical axis e_2 the cumulative distribution F, and $(O, \bar{O})$ the 45-degree line. As long as $\omega > F(\omega)$, whites have an incentive to exit the neighbourhood, causing a reduction of ω, and this process continues until there are no whites left; $\omega = 0$. If, on the other hand, $\omega \leq F(\omega)$, additional whites have an incentive to enter, and this process continues until $\omega = 1$. Thus, the process has three equilibria, $(O, O^*, \bar{O})$, of which the two extreme ones, either only blacks or no blacks in the neighbourhood, are stable, and the mixed one, with ω^* whites in the neighbourhood, where $\omega^* = F(\omega^*)$, unstable. The mixed equilibrium defines the *tipping point*. Individuals' preferences differ widely, but only extreme outcomes *emerge* at the social equilibrium. Schelling (1978) underscores how outcomes that persist may not be what individuals had intended.

Could such a stark outcome be due to the fact that the respective populations of individuals are not being replenished? It turns out that, if one goes deeper and allows for turnover and stochastic shocks, persistence of stable states may be rigorously characterized by means of the tools of stochastic stability theory (Blume and Durlauf 2003; Young 1998). Multiplicity of equilibria allows, of course, for accidents of history to become reinforced over time.

Emergence of Urbanization

The concentrated economic activity that we associate with the emergence of cities punctuates the



Emergence, Fig. 1 Neighbourhood tipping, poverty traps

physical and economic landscape throughout the world. How did it emerge? While small-scale agriculture and home production could be reasonably accurately referred to as spatially uniform distribution of economic activity, the world population is increasingly concentrated in cities. Also, urbanization has been closely associated with economic development.

Let us consider a simple setting where utility U depends on individual productivity, itself an increasing function $f(n_l)$, of the total number of others in the same location, n_l, and on the share of a fixed resource, R. Even when utility is assumed to be increasing and concave in both arguments, it is initially increasing, as a function of n_l, may reach a peak at n^*, and then may start decreasing. In other words, a larger population initially means more innovation and mutually beneficial interaction until congestion offsets them. Consider then two alternative locations, $l = 1, 2$, that do not interact spatially, and a total of N individuals who wish to locate so as to maximize utility. At a locational equilibrium, individuals must be indifferent as to where they locate. If $N < 2n^*$, the symmetric equilibrium, where $n_1 = n_2 = \frac{1}{2}N$, is unstable and agglomeration – that is, either site occupied by the entire population – is stable. Therefore, the trade-off between the value of agglomeration and the cost of congestion moves the economy away from the symmetric outcome (Anas 1992).

Consider next a setting where interactions do explicitly depend on distance to others, as with accessibility to others being valued and congestion disliked. If individuals are allowed to relocate, with probabilities that depend on expected utilities in each site relative to all other sites, then a dynamic model may be formulated that describes locational outcomes for an entire population. The economy may attain steady states that are either uniform (populations are equal across all sites) or uneven (with some sites having large and others small populations). Such a stylized reduced-form model of spatial patterns of human settlements (see Papageorgiou and Smith 1983) yields spatially uniform outcomes that are either stable or unstable. Agglomeration is determined by the interplay between the value of agglomeration and the cost of congestion. If the former dominates, spatially uniform steady states are unstable. Fujita et al. (1999), Chaps. 6 and 17) develop a model with ingredients from economic geography that incorporates trading costs and also allows for uniform distributions of economic activity to exhibit different stability properties. Again, conditions under which agglomerations prevail possess intuitive economic appeal.

E

Emergence of Poverty Traps

In a standard neoclassical growth model that extends over discrete time, with a demographic structure consisting of two overlapping generations and individuals living for two periods, working only in the first and retiring in the second, individual savings would be proportional to the wage rate under Cobb–Douglas preferences. Let the aggregate production function expressing output Y_t as a function of capital, labour and total factor productivity, K_t, L_t, A, respectively, be of the constant elasticity of substitution form,

$$Y_t = A\left(\delta K_t^{1-\frac{1}{\delta}} + (1-\delta)L_t^{1-\frac{1}{\delta}}\right)^{\frac{\sigma}{\sigma-1}}.$$

If the elasticity of substitution is sufficiently small – that is, complementarity between capital and labour is high – and total factor productivity sufficiently large, the time map of the economy – that is, the amount of capital per person next period (axis e_2) as a function of the amount of capital per person in the present period (axis e_1) – may be loosely graphed, as in Fig. 1. Therefore, depending upon the economy's starting point, it may end up at a steady state either with high or with low capital per person at a steady state. The mid-range ('symmetric') steady state is unstable. Therefore, conditions of productive complementarities, (even small) initial differences in capital per person, and possibly historical accidents as well across countries in terms of characteristics and endowments when growth starts, mitigate in favour of an explanation for

inequalities in incomes per person across different countries. The same mechanism worldwide produces sharply different outcomes (see Azariadis and Stachurski 2006, for an in-depth treatment).

Similar arguments may be developed in order to understand persistence in the inequality of the distribution of wealth within an economy. Matsuyama (2006) presents a model of emergent class structure, in which a society inhabited by inherently identical households may, depending upon parameter values, be endogenously split into the rich bourgeoisie and the poor proletariat. For some parameter values, the model has no steady state where all households remain equally wealthy. The model predicts emergent class structure or the rise of class societies. Even if every household starts with the same amount of wealth, the society will experience 'symmetry breaking' and will be polarized into two classes in steady state, where the rich maintain a high level of wealth partly due to the presence of the poor, who have no choice but to work for the rich at a wage rate strictly lower than the 'fair' value of labour.

It is worth noting that similar modelling tools may be used to express Adam Smith's famous dictum that 'the division of labour is limited by the extent of the market' and thus endogenize specialization (Weitzman 1994). The division of labour emerges as individuals in an economy acquire specialized roles.

Emergent Structures in International Economics: Autarky, Specialization, and International Currencies

Krugman (1995) and Matsuyama (1995) discuss how a world economy where all countries are initially identical and live in autarky (a 'symmetric' outcome) leads to a world that is separated into rich and poor regions, once countries engage in international trade. International trade *causes* specialization and agglomeration of different economic activities in different regions of the world to emerge, with some countries being rich and others poor. In several similarly motivated papers, Matsuyama (in particular, Matsuyama 2004, 2006) shows the effects of financial market globalization on the cross-country pattern of development in the world economy. In the absence of the international financial market, the world economy converges to the symmetric steady state, and the cross-country difference disappears in the long run. Financial market globalization causes the instability of the symmetric steady state and generates stable asymmetric steady states, in which the world economy is polarized into the rich and the poor. The world output is smaller, the rich are richer and the poor are poorer in these asymmetric steady states than in the (unstable) symmetric steady state. The model thus demonstrates the possibility that financial market globalization may cause, or at least magnify, inequality among nations, and that the international financial market is a mechanism through which some countries become rich at the expense of others. Furthermore, the poor countries cannot jointly escape from the poverty trap by merely cutting their links to the rich. Nor would foreign aid from the rich to the poor eliminate inequality; as in a game of musical chairs, some countries must be excluded from being rich.

Especially at times of political and economic upheavals, many different national currencies may circulate simultaneously within and across countries. From a modelling viewpoint, such circumstances fit neatly multiplicity of equilibria. Emergence of a particular currency as an international currency, which in turn depends on the degree of economic and financial integration, may be more of a decentralized phenomenon then the emergence and establishment of a national currency (Matsuyama et al. 1993). To start with, a national currency is typically fiat money, whose use is decreed although not necessarily ensured. World monetary history suggests that a bewildering variety of commodities have served as medium of exchange, unit of account and store of value, and may have coexisted at times of financial uncertainties. It has been known at least since Menger (1892) that fiat money comes to dominate other options, thus leading to establishment of monetary equilibria, because individuals accept fiat money in trade when it is convenient and they trust that others will do the same. Such an outcome may be fragile,

when trust in the currency is weakened, especially in time of war and other upheavals. Howitt and Clower (2000) employ 'rules' concerning transactor behaviour (instead of relying on a priori principles of equilibrium and rationality) to show computationally commodity 'money' as a possible emergent property of interactions between gain-seeking transactors who are unaware of any system-wide consequences of their own actions. Similar is the emergence of standards in new industries described by many writers.

Concluding Remarks

The scientific literature, along with popular science literature, on emergence has sought to explain the emergence of persistent patterns as outcomes of dynamic interactions between individuals, groups of individuals and other entities. Such emergence is typically intrinsic to specific nonlinear dynamic processes and represents international currency. Not all possible outcomes may be sustained at equilibrium, and economic and political structures emerge as a result of self-organization. Future research needs to go beyond evolutionary thinking and also deal with emergence in the context of purposeful action by forward-looking agents, as opposed to social outcomes of decentralized interactions of many agents.

See Also

- ▶ Poverty Traps
- ▶ Spontaneous Order

Bibliography

Anas, A. 1992. On the birth and growth of cities: Laissez-faire and planning compared. *Regional Science and Urban Economics* 22: 243–258.

Azariadis, C., and J. Stachurski. 2006. Poverty traps. In *Handbook of economic growth*, ed. P. Aghion and S.N. Durlauf. Amsterdam: North-Holland.

Blume, L.E., and S.N. Durlauf. 2001. The interactions-based approach to socioeconomic behavior. In *Social dynamics*, ed. S.N. Durlauf and H. Peyton Young. Princeton, NJ: Princeton University Press.

Blume, L.E., and S.N. Durlauf. 2003. Equilibrium concepts for social interaction models. *International Game Theory Review* 5: 193–209.

Durlauf, S.N. 1997. Statistical mechanics approaches to socioeconomic behavior. In *The economy as an evolving complex system II*, ed. W.B. Arthur, S.N. Durlauf, and D. Lane. Redwood City, CA: Addison-Wesley.

Erdös, P., and A. Renyi. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17–61.

Fujita, M., P.R. Krugman, and A.J. Venables. 1999. *The spatial economy*. Cambridge, MA: MIT Press.

Holland, J.H. 1998. *Emergence: From chaos to order*. Reading, MA: Addison-Wesley.

Howitt, P., and R. Clower. 2000. The emergence of economic organization. *Journal of Economic Behavior and Organization* 41: 55–84.

Ioannides, Y.M. 1990. Trading uncertainty and market form. *International Economic Review* 31: 619–638.

Kauffman, S.A. 1995. *At home in the universe: The search for the laws of self organization and complexity*. Oxford: Oxford University Press.

Kirman, A.P. 1983. Communication in markets: A suggested approach. *Economic Letters* 12: 1–5.

Kirman, A.P. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6(2): 117–136.

Krugman, P.R. 1995. Complexity and emergent structure in the international economy. In *New directions in trade theory*, ed. A.V. Deardorff, J. Levinsohn, and R.M. Stern. Ann Arbor: University of Michigan Press.

Matsuyama, K. 1995. Comment on P. Krugman, 'Complexity and Emergent Structure in the International Economy'. In *New directions in trade theory*, ed. A.-V. Deardorff, J. Levinsohn, and R.M. Stern. Ann Arbor: University of Michigan Press.

Matsuyama, K. 2004. Financial market globalization, symmetry breaking and endogenous and endogenous inequality of nations. *Econometrica* 72: 853–884.

Matsuyama, K. 2006. The 2005 Lawrence R. Klein lecture: Emergent class structure. *International Economic Review* 47: 327–360.

Matsuyama, K., N. Kiyotaki, and A. Matsui. 1993. Toward a theory of international currency. *Review of Economic Studies* 60: 283–307.

Menger, C. 1892. On the origins of money. *Economic Journal* 2: 239–255.

Newman, M.E.J., S.H. Strogatz, and D.J. Watts. 2001. Random graphs with arbitrary degree distribution and their applications. *Physical Review E* 64: 026118.

Papageorgiou, Y.Y., and R.S. Smith. 1983. Agglomeration as a local instability of spatially uniform steady-states. *Econometrica* 51: 1109–1119.

Schelling, T.C. 1978. *Micromotives and Macrobehavior*. New York: Norton.

Stanford Encyclopedia of Philosophy. 2002. Emergent properties. Online. Available at http://plato.stanford.edu/entries/properties-emergent. Accessed 17 Nov 2006.

Weitzman, M.L. 1994. Monopolistic competition with endogenous specialization. *Review of Economic Studies* 61: 45–56.
Young, H.P. 1998. *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton: Princeton University Press.

Emerging Markets

Joshua Aizenman

Abstract

The club of high-performing emerging markets is fairly concentrated in East Asia. Their TFP growth may not be extraordinary, though their growth rate is unprecedented. Factors argued to promote growth include trade, investment, external financing, and good governance. The importance of external financing is overrated – higher growth induces higher saving rate, allowing investment to be self-financed. Institutional changes as the key for take-off remains debatable – India and China took off without any prior major institutional overhaul. Allowing newcomers to challenge incumbents and the capacity to adjust policies to shocks may be the keys for sustainable growth.

Keywords

Agency problems; Asian miracle; Emerging markets; External financing; Financial liberalization; Financial risk; Growth and governance; Growth and institutions; Growth and international trade; High-performing Asian economies; Moral hazard; Savings; Shocks; Solow, R.; Take-off; Total factor productivity

JEL Classifications

O16

'Emerging markets' are countries or markets that are not well established economically and financially, but are making progress in that direction.

The growing focus on emerging markets follows exciting developments during the second half of the 20th century – the emergence of a growing class of (formerly) poor countries that took off, and managed to close half of their income gap with the OECD countries within a generation or two. Remarkably, from 1960 to 1989 seven high-performing Asian economies (HPAEs) experienced unprecedented growth rates of the real GDP per capita in the range of four to seven per cent. This phenomenon has been the focus of a notable research report by the World Bank (1992), whose title *The East Asian Miracle* suggests a possible, though controversial, interpretation. The big story of recent years has been that the two most populous countries, China and India, joined the HPAE club. With few exceptions (such as Chile and Botswana), the club of high-performing emerging markets is fairly concentrated in East Asia. The HPAEs' remarkable growth rates during recent decades imply a sizable drop in global poverty rates, also entailing greater concentration of the incidence of extreme poverty, mostly in Africa (see Fischer 2003). Yet the emerging markets phenomenon goes well beyond Asia, encompassing a growing share of developing countries that are closing, though at a lower rate than the HPAEs, their income gap with the OECD countries.

These developments were in sharp contrast to the pessimistic predictions made in the 1950–60s by several influential economic growth models (for a review, see Easterly 1999). The HPAE experience dispelled most of these fears. The superior performance of the HPAEs illustrated that the fast growth option is viable, raising pertinent questions, and stirring a lively debate. While the World Bank (1992) dubbed the experience of the HPAEs a 'miracle', Young (1995) questioned this 'miraculous' interpretation, arguing that it is in line with Solow's growth model. Specifically, he reasoned that most of the growth has been the outcome of very high rates of investment in tangible and human capital, and a sizable increase in labour market participation. Controlling for these factors, Young found that the HPAEs' total factor productivity growth is in line with the historical experience of other countries. The debate about

the role of accumulation in accounting for the HPAE experience is not over, yet the large drop of the growth rate of Japan in the 1990s, and the East Asian financial crisis of 1997, somehow deflated the 'East Asian miracle' hypothesis, suggesting the onset of Solow's growth convergence. Even if Young's thesis is correct, the speed and relative smoothness of the convergence of the HPAEs to the OECD's development level are without precedent. It raises questions about the obstacles preventing other countries from accomplishing this task, and about the ways to facilitate the take-off process in other regions.

The HPAE take-offs have been associated with fast growth of exports climbing, over time, the technology ladder of trade. This led to a lively debate about the importance of exports as the engine of growth: is the dominant causal association from exports to growth or vice versa? Earlier studies inferred that trade liberalization enhances growth (Ben-David 1993; Edwards 1998), a point disputed by Rodríguez and Rodrik (2001). Several authors revisited this issue, applying better controls, inferring strong growth effects of trade openness. Frankel and Romer (1999) applied measures of the geographic component of countries' trade to obtain instrumental variables estimates of the effect of trade on income. They inferred that ordinary least square (OLS) estimates understate the effects of trade, and that trade has a significant large positive effect on income. The contrast between the economic performance of the Soviet Union and that of China in the second part of the 20th century suggests another advantage of export orientation: it imposes a powerful market test on domestic output. Since exports must meet the quality and pricing tests of the global market, export-led growth limits potential distortions induced by 'growth promoting' domestic policies. Specifically, it prevents Soviet Union-type superficial economic growth induced by forced investment, growth that may result in inferior products that would be wiped out in the absence of protection. Export-oriented growth also forces countries to move faster towards the technological frontier in order to survive competitive global pressures.

Some of the obstacles preventing countries from taking off arise from political economy factors. Specifically, as growth is frequently associated with the emergence of new sectors and new elites, incumbent policymakers opt to block development in an attempt to preserve their rents and their grip on power. This phenomenon was vividly illustrated at the micro level by De Soto (1989), and was shown to be a major impediment to growth (see Parente and Prescott 2005). As the burden of the low growth would mostly affect future generations, the low growth equilibrium may persist with limited opposition. Proponents of this view point out that free commerce, both internal (between provinces or states in a union) and international, provides a powerful constraint on an incumbent's ability to block development.

The importance of external financing and financial integration in the development process remains a hotly debated topic. Advocates of financial liberalization in the early 1990s argued that external financing would alleviate the scarcity of saving in developing countries, inducing higher investments and thus higher growth rates. In contrast, Rodrik (1998) and Stiglitz (2002) questioned the gains from financial liberalization. Indeed, the 1990s experience with financial liberalization suggests that the gains from external financing are overrated – the bottleneck inhibiting economic growth is less the scarcity of saving and more the scarcity of good governance. This can be illustrated by tracing the patterns of self-financing ratios, measuring the share of tangible capital financed by past national saving (see Aizenman et al. 2004). Higher self-financing rates of the nation's stock of capital are associated with a significant *increase* in growth rates. Remarkably, the wave of financial reforms in the 1990s led to deeper diversification, where greater inflows from the OECD financed comparable outflows from developing countries, with little effect on the availability of resources to finance tangible investment.

These findings are consistent with several interpretations. The first deals with risk: agents in various countries may react to exposure to financial risk differently. The desire to diversify these risks may lead to two-way capital flows,

with little change in net positions (see Dooley 1988). The ultimate obstacles limiting external financing may be related to acute moral hazard and agency problems – sovereign states, decision makers and corporate insiders pursue their own interests at the expense of outside investors (see Gertler and Rogoff 1990; Stulz 2005). An alternative interpretation follows Caroll and Weil (1994), who found that statistical causality runs from higher growth rates to higher saving rates. They conjectured that the growth-saving causality may be explained by habit formation, where consumers' utility depends on both present and past consumption. 'Habit formation', however, may be observationally equivalent to adaptive learning in the presence of uncertainty in countries where private savings are taxed in arbitrary and unpredictable ways, credibility must be acquired as an outcome of a time-consuming learning process. In these circumstances, a higher growth rate provides a positive signal about the competence and the intentions of the administration, increasing saving and investment over time. Consequently, agents in countries characterized by greater political instability and polarization would be more cautious in increasing their saving and investment rates following a reform. Hence, accomplishing take-offs in Latin America may be much harder than in Asia, explaining Latin America's relatively low growth rate. (Various studies pointed out that policy uncertainty and political instability reduce private investment and growth; see Ramey and Ramey 1995; Aizenman and Marion 1999).

I close this review with an outline of open issues. The positive association between the equality of institutions and growth is well documented, yet the precise role of institutions in the development process remains debatable. Acemoglu et al. (2003) inquired how the colonial history of a developing country affects the quality of institutions, concluding that distortionary macroeconomic policies are more likely to be symptoms of underlying institutional problems rather than the main causes of economic volatility. Yet this interpretation does not satisfactorily explain the role of institutions in the growth process. The remarkable take-offs of China and India in recent decades, episodes directly affecting about a third of the global population, cannot obviously be explained by reference to institutional changes. This suggests that there is no simple correspondence or causality between growth and institutions. A tentative answer is provided by Rodrik (1999), who identifies a nonlinear interaction between shocks, polarization of a society and the quality of institutions. This argument suggests the key importance of the capacity of societies to adjust policies to shocks. A deeper understanding of the interaction between history, geography, polarization and institutions remains a challenge awaiting future research.

The exciting developments associated with the emergence of a growing class of (formerly) poor countries that took off implies that the rewards for adopting the proper growth incentives are high. A remaining challenge is how to facilitate the widening of the emerging market club, and how to minimize the prospects of new conflicts associated with the emergence of new economic powers like China and India.

See Also

- ▶ Development Economics
- ▶ Growth and Institutions
- ▶ Growth and International Trade
- ▶ Solow, Robert (Born 1924)

Bibliography

Acemoglu, D., S. Johnson, J. Robinson, and Y. Thaicharoen. 2003. Institutional causes, macroeconomic symptoms: Volatility, crises and growth. *Journal of Monetary Economics* 50: 49–123.

Aizenman, J., and N. Marion. 1999. Volatility and investment: Interpreting evidence from developing countries. *Economica* 66: 157–179.

Aizenman, J., B. Pinto, and A. Radziwill. 2004. Sources for financing domestic capital – is foreign saving a viable option for developing countries? Working

Paper No. 1007. Department of Economics, University of California, Santa Cruz.
Ben-David, D. 1993. Equalizing exchange: Trade liberalization and income convergence. *Quarterly Journal of Economics* 108: 653–679.
Caroll, C., and D. Weil. 1994. Saving and growth: A reinterpretation. *Carnegie Rochester Conference Series* 40: 133–192.
De Soto, H. 1989. *The other path*. New York: Harper and Row.
Dooley, M. 1988. Capital flight: A response to differences in financial risks. *IMF Staff Papers* 35: 422–436.
Easterly, W. 1999. The ghost of financing gap: Testing the growth model used in the international financial institutions. *Journal of Development Economics* 60: 423–438.
Edwards, S. 1998. Openness, productivity and growth: What do we really know? *Economic Journal* 108: 383–398.
Fischer, S. 2003. Globalization and its challenges. *American Economic Review* 93: 1–30.
Frankel, J., and D. Romer. 1999. Does trade cause growth? *American Economic Review* 89: 379–399.
Gertler, M., and K. Rogoff. 1990. North-South lending and endogenous domestic capital market inefficiencies. *Journal of Monetary Economics* 26: 245–266.
Parente, S., and E. Prescott. 2005. A unified theory of the evolution of international income levels. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf, Vol. 1. Amsterdam: North-Holland.
Ramey, G., and V. Ramey. 1995. Cross-country evidence on the link between volatility and growth. *American Economic Review* 85: 1138–1151.
Rodríguez, F. and D. Rodrik. 2001. Trade policy and economic growth: A skeptic's guide to the cross-national evidence, *Macroeconomics Annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press for NBER.
Rodrik, D. 1998. In *Who needs capital-account convertibility? In Should the IMF pursue capital account convertibility? Essays in international finance, No. 207*, ed. P. Kenen. Princeton: Princeton University Press.
Rodrik, D. 1999. Where did all the growth go? External shocks, social conflict, and growth collapses. *Journal of Economic Growth* 4: 358–412.
Stiglitz, J. 2002. *Globalization and its discontents*. New York: W. W. Norton.
Stulz, R. 2005. The limits of financial globalization. *Journal of Finance* 60: 1595–1638.
World Bank. 1992. *The East Asian miracle*. Washington, DC: World Bank.
Young, A. 1995. The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 110: 641–680.

Empirical Likelihood

Yuichi Kitamura

Abstract

Empirical likelihood (EL) is a method for estimation and inference without making distributional assumptions. Viewed as a nonparametric maximum likelihood estimation procedure (NPMLE), it approximates the unknown distribution function with a discrete distribution, then applies the ML estimation method. Alternatively, EL can be regarded as a minimum divergence estimation procedure. EL works well for estimating moment condition models, though it applies to other models as well. The large deviation principle (LDP) and other techniques show that EL has many optimality properties.

Keywords

Blockwise empirical likelihood; Empirical likelihood; Empirical likelihood ratio; Generalized empirical likelihood; Generalized method of moments; Kernel regression technique; Lagrange multiplier; Large deviation principle; Maximum likelihood; Nonparametric maximum likelihood estimation; Semiparametric estimation; Vector autoregressions

JEL Classifications

C14

Introduction

Empirical likelihood (EL) is a method for estimation and inference without making distributional assumptions. The main feature of EL is the use of a discrete distribution to approximate the unknown distribution function nonparametrically, where the approximating discrete distribution is

typically supported by empirical observations. Owen (1988) and subsequent papers considered applications of this approach to moment condition models. Their important discovery is that EL, which can be interpreted as a nonparametric maximum likelihood estimation (NPMLE) method, possesses many desirable asymptotic properties that are analogous to those of parametric likelihood procedures. To describe more details of empirical likelihood, consider i.i.d. data $\{z_i\}_{i=1}^n$, where each z_i is distributed according to an unknown probability distribution F_0. Suppose the expectation of an $\mathbb{R}^q$-valued function $g(z,\theta_0)$, which is known up to the finite-dimensional parameter θ_0 in $\Theta \subset \mathbb{R}^k$, is restricted to be zero:

$$E[g(z,\theta_0)] = \int g(z,\theta_0)\mathrm{d}F_0(z) = 0. \quad (1)$$

Let Δ denote the simplex $\{(p_1,\ldots,p_n) : \sum_{i=1}^n p_i = 1,\ 0 \le p_i, i = 1,\ldots,n\}$. Each vector $(p_1, \ldots, p_n) \in \Delta$ 'parametrizes' the unknown distribution F_0 by $\hat{F}_n(z) = \sum_{i=1}^n p_i 1\{z_i \le z\}$, $z \in \mathbb{R}$ ($1\{\cdot\}$ signifies the usual indicator function). This is the approximating discrete distribution mentioned above. The nonparametric loglikelihood function to be maximized is

$$\ell_{\mathrm{NP}} = \sum_{i=1}^n \log p_i,\ \sum_{i=1}^n g(z_i,\theta)p_i = 0,\ (p_1,\ldots,p_n) \in \Delta, \theta \in \Theta.$$

Let $\left(\hat{\theta}_{\mathrm{EL}}, \hat{p}_{\mathrm{EL1}}, \ldots, \hat{p}_{\mathrm{ELn}}\right)$ denote the value of $(\theta, p_1, \ldots, p_n) \in \Theta \times \Delta$ that maximizes ℓ_{NP}. This is called the (maximum) empirical likelihood estimator. The NPMLE for θ and F are $\hat{\theta}_{\mathrm{EL}}$ and $\hat{F}_{\mathrm{EL}} = \sum_{i=1}^n \hat{p}_{\mathrm{EL}i} 1\{z_i \le z\}$ One might expect that the high dimensionality of the parameter space $\Theta \times \Delta$ makes the above maximization problem intractable for any practical application. Fortunately, that is not the case, if one uses the following nested procedure. First, fix θ at a value in Θ and consider the loglikelihood with the parameters $(p_1, \ldots, p_n)$ 'profiled out':

$$\ell(\theta) = \max \ell_{\mathrm{NP}}(p_1,\ldots,p_n) \text{ subject to } \sum_{i=1}^n p_i = 1,\ \sum_{i=1}^n p_i g(z_i,\theta) = 0. \quad (2)$$

A straightforward application of the Lagrange multiplier method shows that $\ell(\theta)$ is represented by

$$\ell(\theta) = \min_{\gamma \in \mathbb{R}^q} -\sum_{i=1}^n \log(1 + \gamma' g(z_i,\theta)) - n\log n \quad (3)$$

(see, for example, Kitamura 2006). The numerical evaluation of the function $\ell(\cdot)$ is easy, because (3) is a low-dimensional convex maximization problem, for which a simple Newton algorithm works. Second, obtain the empirical likelihood estimator $\hat{\theta}_{\mathrm{EL}}$ as the maximizer of (3). The maximization of $\ell(\theta)$ with respect to θ is typically carried our using a nonlinear optimization algorithm.

Basic properties of the empirical likelihood procedure are now well-understood. The EL estimator $\hat{\theta}_{\mathrm{EL}}$ is $n^{1/2}$-consistent and asymptotically normal. Let D and S denote $E[\nabla_\theta g(z, \theta_0)]$ and $E[g(z,\theta_0)g(z,\theta_0)']$, then its asymptotic distribution is given by $\mathrm{N}(0,(D'SD)^{-1})$. Also, suppose R is a known $\mathbb{R}^s$-valued function of θ, and the econometrician poses a hypothesis that θ_0 is restricted as $R(\theta_0) = 0$, where the s restrictions are independent. This can be tested by forming a nonparametric analogue of the parametric likelihood ratio statistic. Let $r = -2(\sup_{\theta : R(\theta) = 0}\ell(\theta) - \sup_{\theta \in \Theta}\ell)$, then this obeys the chi-square distribution with s degrees of freedom asymptotically under the null. The factor r is called the empirical likelihood ratio (ELR) statistic. ELR also applies to testing overidentifying restrictions: see section "EL and the Large Deviation Principle". These properties and other basics of EL and related methods have been studied extensively in the literature (see Qin and Lawless 1994; Imbens 1997; Kitamura 1997; Kitamura and Stutzer 1997; Smith 1997; Imbens et al. 1998; Newey and Smith 2004).

An alternative way to motivate EL is to use a minimum divergence estimation framework. Let f and g denote the density functions or the probability functions of distribution functions F and G. Define a 'divergence measure' between F and G to be

$$D(F,G) = \int \varphi\left(\frac{f(z)}{g(z)}\right) g(z)\mathrm{d}z, \tag{4}$$

for a convex function φ. It is easy to see that $D(\ ,G)$ is minimized at G. Let

$$\mathcal{F}(\theta) = \left\{F : \int g(z,\theta)\mathrm{d}F = 0, F \text{ is a CDF}\right\}$$

Then $\mathcal{F} = \cup_{\theta\in\Theta}\mathcal{F}(\theta)$ is the set of all probability distributions that are compatible with the moment restriction (1). Now consider the problem of minimizing the divergence $D(F,F_0)$ with respect to $F \in \mathcal{F}$. In other words, a distribution that is 'closest' to the true distribution F_0 in the class of distributions $\mathcal{F}$ is sought. Pick a value $\theta \in \Theta$ and define

$$v(\theta) = \inf_F D(F,F_0) \text{subject to} \int g(z,\theta) f\mathrm{d}z = 0, \int f\mathrm{d}z = 1. \tag{P}$$

The value $v(\theta)$ is regarded as the minimum divergence between F_0 and the set of distributions that satisfy the moment restriction with respect to $g(z,\theta)$. The nonnegativity of f is maintained if φ is modified so that $\varphi(z) = \infty$ for $z < 0$ (see Borwein and Lewis 1991). The primal problem (P) has a dual problem

$$v^*(\theta) = \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q}\left[\lambda - \int \varphi^*(\lambda + \gamma' g(z,\theta))\mathrm{d}F_0(z)\right], \tag{DP}$$

where φ^* is the convex conjugate (or the Legendre transformation) of φ, that is $\varphi^*(y) = \sup_x[xy - \varphi(x)]$. (DP) is a finite-dimensional unconstrained convex maximization problem. The Fenchel duality theorem implies that $v(\theta) = v^*(\theta)$. Since the true value θ_0 minimizes $v(\theta)$ over Θ, it follows that

$$\theta_0 = \mathrm{argmin}_{\theta\in\Theta} v^*(\theta). \tag{5}$$

Note that the integral in the definition of v^* is the expected value of $\varphi^*(\lambda + \gamma' g(z,\theta))$ with respect the true distribution F_0, which is unknown in practice. A feasible procedure is obtained by replacing the expectation with the sample average, that is

$$\hat{v}(\theta) = \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q}\left[\lambda - \frac{1}{n}\sum_{i=1}^n \varphi*(\lambda + \gamma' g(z_i,\theta))\right]. \tag{6}$$

Corresponding to (5), an appropriate minimum distance estimator takes the form

$$\hat{\theta} = \mathrm{argmin}_{\theta\in\Theta}\hat{v}^*(\theta).$$

This minimum divergence framework yields empirical likelihood as a special case with $\varphi(x) = -\log(x)$ (or equivalently, $\varphi^*(x) = -1 - \log(-y)$). Other choices for φ are, of course, possible. For example, $\varphi(x) = x\log(x)$ yields the 'exponential tilt' estimator (Kitamura and Stutzer 1997), while $\varphi(z) = \frac{1}{2}(x^2 - 1)$ corresponds to the continuous updating GMM estimator (CUE) (Hansen et al. 1996). A convenient parametric family of convex functions known as the Cressie–Read family (Read and Cressie 1988) subsumes these three important cases. If φ belongs to the Cressie–Read family, one can show that the minimum divergence estimator can be written as

$$\hat{\theta} = \mathrm{argmin}_{\theta\in\Theta}\max_{\gamma\in\mathbb{R}^q}\left[\frac{1}{n}\sum_{i=1}^n \kappa(\gamma' g(z_i,\theta))\right] \tag{7}$$

where $\kappa(y) = -\varphi(y+1)$. This is essentially equivalent to the generalized empirical likelihood (GEL) estimator by Smith (1997). Smith (2004) provides a detailed account for GEL.

E

EL and the Large Deviation Principle

Like the conventional asymptotic method, the large deviation principle (LDP) offers first order approximations for various estimators and tests. Unlike the conventional theory, which produces local linear approximations, the LDP provides global nonlinear approximations. It is the latter feature that enables the LDP to yield results not obtained by the conventional linear approximations. For example, the LDP shows that EL enjoys many optimality properties that are not shared by, for example, the conventional GMM estimator.

To introduce the concept of the LDP in the context of moment condition models, suppose the econometrician observes i.i.d. data $(z_1, \ldots, z_n)$, where z_i satisfies the restriction (1). Let A_n be an event as a result of estimation or testing: for example, if one uses an estimator θ_n to estimate θ_0, one may consider $A_n = 1\{\|\theta_n - \theta_0\| > c\}$ for a constant c. Then $\Pr\{A_n\}$ is the probability of the estimator missing the true value by a margin larger than c. Or, in testing a null hypothesis H_0, A_n can represent the event that H_0 is accepted. If the null is incorrect, $\Pr\{A_n\}$ is the probability of type II errors. In either way, $\lim_{n \to \infty} \Pr\{A_n\} = 0$ if the estimator or the test is consistent. The LDP also deals with asymptotic properties, but it is concerned with the limit of the form $\lim_{n\to\infty} \frac{1}{n} \log \Pr\{A_n\}$. (If the limit does not exist, one needs to consider lim inf or lim sup, depending on the purpose of analysis.) Let $-d \leq 0$ denote the above limit so that $\Pr\{A_n\} \approx e^{-nd}$, which characterizes how fast $\Pr\{A_n\}$ decays. The goal is to obtain a procedure that maximizes the speed of decay d.

Kitamura and Otsu (2005) study the estimation of models of the form (1) using the LDP. One complication in the application of the LDP to an estimation problem in general is that an estimator that maximizes the limiting decay rate d with $A_n = 1\{\|\theta_n - \theta_0\| > c\}$ uniformly in unknown parameters does not exist in general, unless the model belongs to the exponential family. A possible way around this issue is to pursue minimax optimality, rather uniform optimality. See Puhalskii and Spokoiny (1998) for a general discussion on such a minimax framework. Note that the probability of the event $A_n = 1\left\{\left\|\hat{\theta}_n - \theta_0\right\| > c\right\}$ depends on θ_0 and F_0, therefore the worst case scenario is given by the pair (allowed in the model (1)) that maximizes $\Pr\{A_n\}$. Suppose an estimator θ_n minimizes this worst-case probability, thereby achieving minimaxity. The limit inferior of the minimax probability provides an asymptotic minimax criterion. Kitamura and Otsu (2005) show that an estimator that attains the lower bound of the asymptotic minimax criterion can be obtained from the EL objective function $\ell(\theta)$ in (2) as follows:

$$\hat{\theta}_{\mathrm{ld}} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, Q_n(\theta), Q_n(\theta) = \sup_{\theta* \in \Theta: \|\theta* - \theta\| > c} \ell(\theta*).$$

Calculating $\hat{\theta}_{\mathrm{ld}}$ in practice is straightforward. If the dimension of θ is high, it is also possible to focus on a low-dimensional sub-vector of θ and obtain a large deviation minimax estimator for it, treating the rest as nuisance parameters.

Kitamura (2001) shows that empirical likelihood dominates other methods in terms of the LDP when applied to overidentifying restrictions testing. Researchers routinely test overidentifying restrictions of the form

$$\int g(z, \theta) \mathrm{d}F = 0 \text{ for some } \theta \in \Theta \text{ and}$$
$$\text{for some distribution function } F, \qquad \text{(O)}$$

with $\dim(\Theta) = k$ and $g \in \mathbb{R}^q$, $q > k$. The log empirical likelihood under the restriction (O) is $\sup_{\theta \in \Theta} \ell(\theta)$; without the restriction, it is $-n \log n$. The ELR test statistic for (O) is the difference of the two multiplied by -2. It is asymptotically distributed according to the χ^2 distribution with $q - k$ degrees of freedom under (O) (Qin and Lawless 1994). Using the notation in the previous section, rewrite the above null in an equivalent form: $(\mathrm{O})' : F_0 \in \mathcal{F}$. It turns out that ELR for $(\mathrm{O})'$ has a property of being uniformly most powerful in an LDP criterion. To state this optimality property of ELR formally, let F denote the set of all probability distribution functions. Practically all reasonable tests for (O) (or $(\mathrm{O})'$) can be represented by a partition $\Omega = (\Omega_1; \Omega_2)$ of F,

such that if the empirical distribution function F_n falls into Ω_1 (Ω_2) one rejects (accepts) (O). It is a straightforward exercise to show that the ELR test rejects the null if the Kullback–Leibler divergence $K(F_n,G)$ between F_n and G, minimized over $G \in \mathcal{F}$, is too large. Therefore ELR is represented by the following partition of F : $\Lambda = (\Lambda_1; \Lambda_2)$, $\Lambda_1 = \{F : \inf_{G \in F} K(F,G) < \eta\}, \Lambda_2 = \Lambda_2 = \Lambda_1^c$ for a positive number η. Following Owen (2001), for an event A_n that involves observations $z_1,\ldots, z_n$ that are randomly sampled from F, let $\Pr\{A_n; F\}$ denote the probability of the event. By applying a mathematical result called Sanov's theorem, it can be shown that

$$\sup_{F* \in \mathcal{F}} \limsup_{n \to \infty} \frac{1}{n} \log \Pr\{F_n \in \Lambda_2; F*\} \leq -\eta.$$

Kitamura (2001) also shows that if the following inequality holds for a test $\Omega = (\Omega_1; \Omega_2)$ that satisfies some regularity conditions (see Kitamura 2001, for the regularity conditions):

$$\sup_{F* \in \mathcal{F}} \limsup_{n \to \infty} \frac{1}{n} \log \Pr\{F_n \in \Omega_2; F*\} \leq \eta,$$

then it must be that

$$\limsup_{n \to \infty} \frac{1}{n} \log \Pr\{F_n \in | \, \Omega_1; F**\} \geq \limsup_{n \to \infty} \frac{1}{n} \log \Pr\{F_n \in \Lambda_1; F**\}$$

for every $F** \notin \mathcal{F}$. The first two of the above three inequalities mean that the ELR test Λ and the arbitrary regular test Ω are comparable in terms of its LDP property of type I error probabilities. But the third inequality implies that the ELR test is no less powerful than the arbitrary test if the LDP of type II error probabilities are used to measure the asymptotic powers of the tests. Note that the third inequality holds for every $F** \notin \mathcal{F}$: that is, it holds uniformly over alternatives. Since the test (Ω_1, Ω_2) is arbitrary, this shows that ELR is uniformly most powerful in an LDP sense. Such a property is sometimes referred to as the Generalized Neyman–Pearson (GNP) optimality.

Higher-Order Asymptotics

An alternative way to see why EL works well is to analyse it using higher-order asymptotics. Newey and Smith (2004) investigate higher-order properties of the GEL family of estimators. To illustrate their findings, it is instructive to look at the first-order condition that the EL estimator satisfies, that is $\nabla_\theta \ell\left(\hat{\theta}_{\mathrm{EL}}\right) = 0$. A straightforward calculation shows that this condition, using the notation $\hat{D}(\theta) = \sum_{i=1}^n \hat{p}_{\mathrm{ELi}} \nabla_\theta g(z_i, \theta)$ and $\hat{S}(\theta) = \sum_{i=1}^n \hat{p}_{\mathrm{ELi}} g(z_i, \theta) g(z_i, \theta)'$, can be written as

$$\hat{D}\left(\hat{\theta}_{\mathrm{EL}}\right)' \hat{S}^{-1}\left(\hat{\theta}_{\mathrm{EL}}\right) \overline{g}\left(\hat{\theta}_{\mathrm{EL}}\right) = 0; \qquad (8)$$

see Theorem 2.3 of Newey and Smith (2004). The factor $\hat{D}\left(\hat{\theta}_{\mathrm{EL}}\right)' \hat{S}^{-1}\left(\hat{\theta}_{\mathrm{EL}}\right)$ can be interpreted as a feasible version of the optimal weight for the sample moment $\overline{g}(\theta) = \frac{1}{n} \sum_{i=1}^n g(z_i, \theta)$. Equation (8) is similar to the first-order condition for GMM, though there are important differences. Notice that the Jacobian term D and the variance term S are estimated by $\hat{D}\left(\hat{\theta}_{\mathrm{EL}}\right)$ and $\hat{S}\left(\hat{\theta}_{\mathrm{EL}}\right)$ in (8). It can be shown that these are semiparametrically efficient estimators of D and S under the moment restriction (1). This means that they are asymptotically uncorrelated with $\overline{g}(\theta_0)$, removing the important source of the second-order bias of GMM. Moreover, the EL estimator does not involve a preliminary estimator, thereby eliminating another source of the second-order bias in GMM. Newey and Smith (2004) formalize this intuition and obtain an important conclusion that the second-order bias of the EL estimator is equal to that of the infeasible method-of-moments estimator that optimally weights $\overline{g}$ by the unknown factor $D'S^{-1}$. In contrast, the first-order condition of GMM takes a similar form, but the terms that correspond to D and S are inefficiently estimated, causing bias. Newey and Smith (2004) note that the first-order conditions of GEL estimators have a form where D is efficiently estimated but S is not, leaving a source of bias that is not present for EL.

Higher-order properties of ELR tests have been studied in the literature as well. One of the significant findings in the early literature of empirical

likelihood is the Bartlett correctability of the empirical likelihood ratio test, discovered by DiCiccio et al. (1991). Consider the ELR test statistic for H_0:$\theta = \theta_0$ in the model (1) with $q = k$. DiCiccio et al. (1991) show that the accuracy of the χ^2 asymptotic approximation for the distribution of the ELR statistic can be improved from the rate n^{-1} to the much faster rate n^{-2} by multiplying it by a factor called the Bartlett coefficient.

Some Variations of EL

EL is applicable to many problems other than (1), but they sometimes require extending and modifying the standard EL method described so far. For example, suppose economic theory implies that the conditional mean of $g(z,\theta_0)$ given a vector of covariates x is zero:

$$E[g(z,\theta_0)|x] = 0 \tag{9}$$

This restriction is stronger than (1). Though one can choose an arbitrary function $a(x)$ of x as an instrument, this can be problematic since (*a*) choosing an instrument that delivers strong identification may be a difficult task, and (*b*) an arbitrary instrument does not achieve efficiency in general. Kitamura et al. (2004) use the kernel regression technique to incorporate the information in the conditional moment restriction into empirical likelihood. Their estimator achieves the semiparametric efficiency bound of the model (9) under weak regularity conditions. While there exist estimators that achieve efficiency in the model, the EL-based estimator has an advantage that finding a preliminary estimator that is consistent is not necessary. A simulation study in Kitamura et al. (2004) indicates that the conditional EL estimator and tests based on it work remarkably well in finite samples. Donald et al. (2003) propose an alternative estimator for (9). Their idea is to use a sequence functions of x as a vector of instruments, then apply EL to the resulting unconditional moment restriction model. By letting the dimension of the instrument vector grow with the sample size in such a way that it spans the 'optimal instrument' asymptotically, their procedure also achieves the semiparametric efficiency bound.

A topic that is closely related to the above is nonparametric specification testing. Suppose, for example, one is interested in testing the specification of a parametric regression model $E[y|x] = m(x, \theta_0)$, where m is parametrized by a vector $\theta_0 \in \Theta$. The null hypothesis of correct specification can be written in terms of a conditional moment restriction for the function $g(z, \theta) = y - m(x, \theta)$; $z = (x', y)'$:

$$E[g(z,\theta)|x] = 0 \text{ for some } \theta \in \Theta \tag{C}$$

Tripathi and Kitamura (2003) shows that a conditional version of the ELR test applies to the above problem. They propose a simple procedure: reject (C) if the maximized value of the conditional empirical likelihood function, which is essentially the one used in Kitamura et al. (2004), is too small. They also calculate the asymptotic power of their test. Their analysis shows that the EL-based testing procedure has an asymptotic optimality property in terms of an average power criterion.

Another example in which EL needs an appropriate modification is a time series model. Suppose the researcher observes a strictly stationary and weakly dependent time series $\{z_1, \ldots, z_t\}$, and each z_t satisfies the moment condition $E[g(z_t, \theta_0)] = 0$, $\theta_0 \in \Theta$. Applying EL to this model ignoring dependence is inappropriate; it leads to efficiency loss, and the chi-square asymptotics of the ELR test break down.

There are at least three alternative ways to deal with the problem caused by dependence. The first approach is to parametrize the dynamics using a reduced form time series model such as a vector autoregression (VAR) model (Kitamura 2006). While straightforward, this approach involves the risk of mis-specifying the dynamics, and reduces the appeal of EL as nonparametric likelihood. The second approach is the blocking method proposed by Kitamura and Stutzer (1997) and Kitamura (1997). The idea is to form data blocks by taking consecutive observations, and apply EL to them. This is termed blockwise

empirical likelihood (BEL). BEL preserves the dependence information in the data, in a fully nonparametric manner. The third approach is a hybrid of the first and the second approaches (Kitamura 2006). That is, one applies a low order parametric filter to lessen the degree of dependence in the data, then apply BEL to the filtered data. While this does not change the desirable asymptotic property of BEL, it appears to have advantages in finite samples when applied to a time series that is highly persistent.

See Also

- ▶ Generalized Method of Moments Estimation
- ▶ Semiparametric Estimation
- ▶ Vector Autoregressions

Bibliography

Borwein, J.M., and A.S. Lewis. 1991. Duality relationships for entropy-type minimization problems. *SIAM Journal of Control and Optimization* 29: 325–338.

DiCiccio, T., P. Hall, and J. Romano. 1991. Empirical likelihood is Bartlettcorrectable. *Annals of Statistics* 19: 1053–1061.

Donald, S.G., G.W. Imbens, and W.K. Newey. 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* 117: 55–93.

Hansen, L.P., J. Heaton, and A. Yaron. 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* 14: 262–280.

Imbens, G.W. 1997. One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies* 64: 359–383.

Imbens, G.W., R.H. Spady, and P. Johnson. 1998. Information theoretic approaches to inference in moment condition models. *Econometrica* 66: 333–357.

Kitamura, Y. 1997. Empirical likelihood methods with weakly dependent processes. *Annals of Statistics* 25: 2084–2102.

Kitamura, Y. 2001. Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* 69: 1661–1672.

Kitamura, Y. 2006. Empirical likelihood methods in econometrics: theory and practice. In *Advances in economics and econometrics: Theory and applications, ninth world congress*, ed. R. Blundell, W.K. Newey, and T. Persson. Cambridge: Cambridge University Press.

Kitamura, Y., and T. Otsu. 2005. Minimax estimation and testing for moment condition models via large deviations. Manuscript, Department of Economics, Yale University.

Kitamura, Y., and M. Stutzer. 1997. An information theoretic alternative to generalized method of moments estimation. *Econometrica* 65: 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn. 2004. Empirical likelihood based inference in conditional moment restriction models. *Econometrica* 72: 1667–1714.

Newey, W.K., and R.J. Smith. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72: 219–255.

Owen, A. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75: 237–249.

Owen, A. 2001. *Empirical likelihood*. New York: Chapman and Hall/CRC.

Puhalskii, A., and V. Spokoiny. 1998. On large-deviation efficiency in statistical inference. *Bernoulli* 4: 203–272.

Qin, J., and J. Lawless. 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* 22: 300–325.

Read, T.R.C., and N.A.C. Cressie. 1988. *Goodness-of-fit statistics for discrete multivariate data*. Berlin: Springer.

Smith, R.J. 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal* 107: 503–519.

Smith, R.J. 2004. *GEL criteria for moment condition models*. Working paper: University of Warwick.

Tripathi, G., and Y. Kitamura. 2003. Testing conditional moment restrictions. *Annals of Statistics* 31: 2059–2095.

Employment, Theories of

M. Bronfenbrenner

Theories of employment are actually concerned with involuntary unemployment. They deal with the definition, nature, and causes of such unemployment, and also with economic polices to reduce or alleviate it. They consider such questions as:

How serious is the problem of involuntary unemployment, both in the short and in the longer run?

Is such unemployment a feature of economic equilibrium, or is it an exclusively disequilibrium phenomenon?

Is there a 'natural' or 'normal' or 'non-inflation-accelerating' unemployment rate?

Is there a trade-off between unemployment and inflation rates? If so, what are the terms of trade-off, and are they stable?

Under what circumstances, if any, can assurance of long-term high employment be combined with assurance of price-level stability (or non-accelerating inflation)?

Opposed Positions

Two basic and opposed positions of economists on employment theory may be summarized as follows:

(1) Applying standard supply-and-demand analysis to labour markets makes unemployment a disequilibrium phenomenon, resulting from the prevalence and persistence of real and money wage rates higher than the demand for labour will support. Its solution is the lowering of real wage rates to market-clearing levels, rather than any arbitrary removal of certain classes of workers from the labour supply. Public support for the unemployed may perhaps subsidize search for desirable jobs, but it should not subsidize withdrawal from the labour force. Intentional stimulation of labour demand, as by monetary expansion or fiscal deficits, is apt to kindle or accelerate inflation, and/or to raise interest rates and discourage investment. Limitation of labour-saving technical progress will slow economic growth at the expense of future generations.

(2) Unemployment results from equilibrium between aggregate supply and demand for the national output at a level too low to require the productive services of the full labour supply. The appropriate remedy is expansion of demand by fiscal and monetary measures – increased public spending, lower taxes, accelerated monetary growth, lower interest rates. Removal of particular groups from the labour market – youth, the elderly, secondary workers in families with employed breadwinners -and moratoria on labour-saving innovations may be legitimate devices to reduce unemployment in the short run, as may the export of unemployment by export subsidies and protection against imports.

Variants of the first of these positions – sometimes called neoclassical, but actually much older than the 'neoclassical revolution' of the 1870s – dominated economic orthodoxy in Western countries prior to the 1930s. Variants of the second position, articulated by John Maynard Keynes's General Theory of Employment Interest and Money (1936, esp. ch. 19) are called Keynesian, although none of the basic ideas was precisely new in 1936. (Marx, for example, had gone far beyond Keynes in regarding a 'reserve army of the unemployed' as highly functional in capitalism, its function being to hold wage rates at an established subsistence level.) But much as persistent depression unemployment threatened neoclassicism, persistent and accelerating inflation has later threatened Keynesianism. Revolts against Keynesian neoorthodoxy have taken two opposite tacks; towards reformulated neoclassicism on one side, and towards 'incomes policies' of employment guarantees (with regulated prices and usually also wages) at the opposite end of the spectrum. The current (mid–1980s) situation is variously described as fluid, as chaotic, and as 'in shambles'. It cannot be called 'cut and dried'!

Conflict between neoclassicals and Keynesians is exacerbated by denunciation in each group of the other's policy proposals as dangerously harmful. To the confirmed neoclassicist, artificial demand stimulus, repeated and anticipated, soon raises society's (unofficial) 'discomfort index' by raising the inflation rate more than it lowers the measured unemployment rate. To the confirmed Keynesian, the immediate effect of any real or money wage cut is to deepen recession by shifting purchasing power from 'spenders' (the working class) to 'savers' (capitalists and corporate treasuries).

Neoclassical Employment Theory

Mature neoclassical employment theory, as represented by A.C. Pigou's *Theory of Unemployment* (1933), draws its analysis from Alfred

Marshall's *Principles of Economics* (1890). There had been little formal employment theory in Marshall himself, the most nearly relevant materials being the treatment of derived demand (Book V, ch. vi) with reference to the building trades, rather than the 'wages' chapters of Book VI. Book V includes Marshall's famous 'four laws' governing the extent to which a rise in the demand and price of an output (houses) causes a rise in the demand and wage of an input (building workers). In today's economic terminology and Marshall's order, these laws state that the rise in demand for the input will increase more, the lower the elasticity of substitution between that input and other inputs, the lower the elasticity of demand for the output, the less the importance of that input in the production process for the output, and the less the elasticities of supply of substitute inputs. This seems far removed from employment theory, but Pigou, in successive editions of his *Economics of Welfare* (1920, 1st edition entitled *Wealth and Welfare*, 1912) restated Marshall's laws as conditions under which workers might obtain higher wages, presumably after union organization, with minimum losses of employment. A full mathematical statement, combining all four laws in a single formula for the elasticity of the derived demand for a labour input, dates from J.R. Hicks's *Theory of Wages* (1932). Denoting by E, η, σ, e the respective elasticities of labour demand, output demand, substitution between labour and 'capital', and supply of 'capital', the Hicks equation is:

$$E = \frac{\sigma(\eta + e) + k\sigma(\eta - e)}{\sigma(\eta + e) + k(\eta - e)}$$

with k representing the relative importance of labour in production as measured by the proportionate share of wage payments in total cost. (The equation ignores shifts of consumer demand between more and less labour-intensive commodities.) From Hicks's equation, Marshall's laws follow immediately, with the possible exception of the third one on 'the importance of being unimportant'.

We come now to Pigou's *Theory of Unemployment*. Based on the Marshallian structure and appearing in mid-Depression, this volume is remembered chiefly as the fuse that lit Keynes's *General Theory*, but deserves a better fate. Much of it can be interpreted as standing Pigou's earlier argument on its head, so as to provide us an exposition of conditions under which employment can be *restored* most rapidly in a depression, and with *minimum* cuts in real wages. These conditions are embodied in devices to *raise* the elasticity of demand for labour. These several devices involve shifts in aggregate demand, private and (especially) public from what Pigou calls 'centres' where (1) σ is low to others where it is high, (2) from centres where η is low to others where it is high, (3) from centres where k is low to others where it is high – surely the most important quantitatively – and finally (4) from centres where e is low to others where it is high.

On the more aggregative plane, Pigou seems wavering and inconsistent in the light of fifty years' additional development of macroeconomic theory. Over the long term, he argues in the second chapter of Part V, increases in aggregate demand serve only to raise prices and wages without increasing employment. In the short run, however, they can be helpful if not carried too far. This apparently implies that Say's Law is valid as a long-run proposition, but inoperative in the short run.

Pigou's approach may seem naïve in transferring microeconomic analysis to the macroeconomic plane, and in its failure to examine Say's Law more intensively than it does. But it is far from the labour-and union-bashing that neoclassical economic argumentation is often supposed to represent.

Keynesian Employment Theory

Despite its title, Lord Keynes's *General Theory* is a treatise on the macroeconomic theory of income determination. Its employment theory is confined largely to attacks on the Marshall-Pigou tradition, under the assumption (too obvious to require either stress or detailed development) that aggregate real income and the unemployment rate are inversely related under any given state of technology.

Strengthening and specifying of the Keynesian relation between income on the one hand, and employment and unemployment on the other, came only after advances in econometrics and in computer technology. A standard specification has been contributed by Arthur Okun. 'Okun's Law' or the 'Okun curve', as it is variously known, may be written in disguised differential-equation form. For example, let us denote by $\hat{U}$ the percentage change in the measured unemployment rate over a given period and by $\hat{Y}$ the percentage change in a real-income measure like deflated GNP or GDP over the same period.

$$-\hat{U} = a\left(\hat{Y} - \overline{Y}_0\right) \quad \text{or} \quad \hat{U} + a\left(\overline{Y} - \hat{Y}_0\right) = 0$$

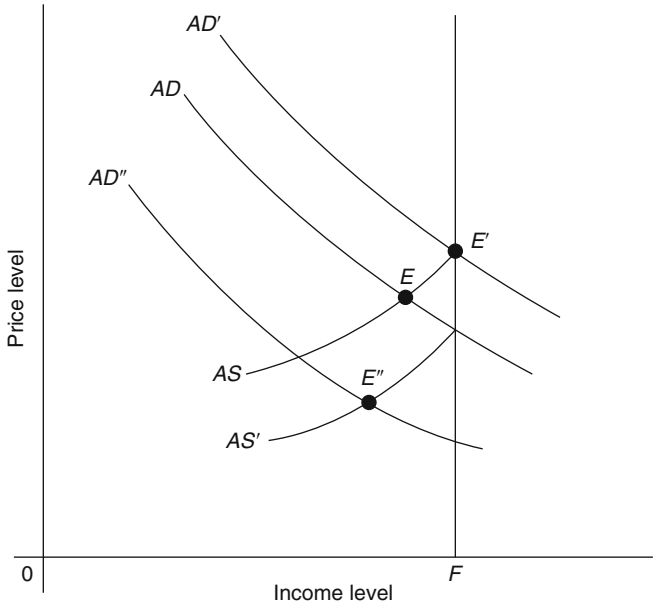
Here a is a statistical parameter, while $\overline{Y}_0$ is the income growth rate estimated to be required if the unemployment rate is not to rise. Neither real nor nominal wage rate movements are taken into explicit account. (The estimates of $\overline{Y}_0$ are based on regressions of high-employment points only, and the slope of this regression is called the country's potential growth rate. The area between such a regression and the country's actual growth path is sometimes called an 'Okun gap'.)

The $\left(a, \overline{Y}_0\right)$ values may of course vary widely, both across countries and over time. For the US 1949–60, Okun's estimate of a was about 0.3, and his estimate of $\overline{Y}_0$ about 3.75 per cent. The usefulness of the widely used Okun analysis is questionable, however, in the presence of supply shocks, wage 'explosions', and similar disturbances. (A productivity jump, for example, might be expected to lower a and raise $\overline{Y}_0$. The combination of these shifts increases the income growth rate required to reduce an existing unemployment rate or to keep that rate from rising.)

Shifting from the empirical to the analytical side, a fruitful development of Keynesian unemployment theory was the distinction between high and low full employment, introduced by A.P. Lerner's *Economics of Employment* (1951, ch. 13). By 'low full employment' Lerner meant essentially what was later called NAIRU in the United Kingdom, namely the non-inflation-accelerating rate of unemployment. By 'high full employment', on the other hand, Lerner meant 100 per cent of the labour force, minus only the frictional lacunae consequent upon job changing. Lerner set low rather than high full employment as the preferred target of employment policy, and outlined a detailed (but probably impractical) scheme of wage controls for attaining and maintaining it. (Most professed Keynesians, more ambitious than Lerner, would strive for high full employment.)

The 'expository Keynesianism' of the textbooks stresses primarily the *shapes* and *parameters* of such relations as the consumption, marginal efficiency, and liquidity functions which determine the level of income in the Keynesian scheme. At a more advanced level, the emphasis appears to be changing, to stress rather the *volatility* of these functions as expectations fluctuate. Leijonhufvud (1968) calls the newer view 'the economics of Keynes' as distinguished from the 'Keynesian economics' of the textbooks. The change in emphasis may also result from the defence of Keynesian macroeconomics, with its employment-theory appendage, against its critics, whose argument can be paraphrased: 'The aggregate demand for the national output depends inversely on the general price level, as the demand for a single commodity depends upon its price. If human wants are insatiable at a zero price level, it follows that there exists a positive price (and likewise wage) level at which full-employment output can be absorbed.' (The critics could not, however, prove that the market-clearing set of full-employment wage rates was at or above 'subsistence', however defined.)

The Keynesian rebuttal, due largely to Clower (1965), introduced the distinction between actual and 'notional' demands for output and especially for labour. In a recession, the demand function for output is weaker than the 'notional' full-employment demand function would be. At the same time, the actual demand for labour is less than the notional one which would prevail were potential employers reasonably sure of selling full-employment output at profitable prices. The impasse or vicious circle could be broken when the demand for output (see Fig. 1) could rise from



Employment, Theories of, Fig. 1 Investment and saving in a 2-period model

AD to *AD′* by public policies which improved the state of confidence – without reference to the multiplier mechanism of the Keynesian textbooks. The equilibrium position could move from point *E*, possibly all the way to point *E′* at the full-employment income level *F*, with money wages remaining the same and with no need to press aggregate supply *AS* vertically downward to *AS′* as by a wage cut.

Now suppose, in the same recession, there was to be a complete 'hands-off' policy. As unemployment continued with nothing done about it beyond calls for wage reductions, notional aggregate demand would fall, perhaps as far as *AD″* if the policy were to lead to budget-balancing and monetary contraction. Even if hard times and wage concessions from labour eventually forced aggregate supply to the *AS′* position, the result would be hyper-deflation rather than recovery. (The Hoover debacle of 1931–2 in the United States and the contemporaneous Brüning one in Germany are cases in point.) By the time aggregate supply had reached *AS′*, the equilibrium point *E″* would prevail, with output (and therefore employment) below those at *E*, even though deflationary cost-cutting would have restored high employment had aggregate demand remained at *AD*.

Equilibrium is then not unique. Depending upon the state of notional as well as actual demand, there are an infinite number of possible equilibria at as many levels of income and employment. We cannot be sure *a priori* that wage deflation will produce increases in employment.

Conclusion

The British Broadcasting Company featured in December 1944 a series of postwar-planning programmes entitled *Jobs for All*. These were inspired not only by Lord Keynes's *General Theory* but by the extensions and applications proposed by Michal Kalečki in the Oxford Institute of Statistics' *Economics of Full Employment* (Burchardt et al. 1944). (Kalečki, who would later assume temporary leadership of British Keynesianism after Keynes's own death in 1946, had proposed 'incomes policies' and income redistribution as preferable to either deficit finance or stimulation of private investment as a route to full employment.) A junior member of that Oxford team, and a speaker in the BBC *Jobs for All* series, was G. D. N. Worswick. In July 1984,

the same G.D.N. Worswick, now Professor at Oxford and President of the Royal Economic Society, delivered his presidential address on 'Jobs for All?', later published in the *Economic Journal* (Worswick 1985). It was the same subject, but note the question mark.

The substance of Worswick's address was that, unfortunately, that question mark belonged in his new title, and could not be expunged even after forty additional years of planning, theorizing, and experimentation. We quote from his final paragraph (p. 14):

> When it comes to action, *[The Economics of Full Employment]* was already too optimistic. We assumed that trade unions would readily accept some limitations on free collective bargaining as a small price to pay for ending unemployment. There was too little recognition that it is my restraint which is necessary to secure your employment. Is it possible to devise schemes which are not only of advantage for the national economy, or for workers as a whole, but can also be seen to be to the advantage . . . of members of trade unions who are already in employment? This is a task for the new generation of economists to undertake. Until a lasting solution is found, the question mark after my title must remain.

See Also

- ▶ Aggregate Demand and Supply Analysis
- ▶ Effective Demand
- ▶ Full Employment
- ▶ Involuntary Unemployment
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Keynes's General Theory
- ▶ Output and Employment
- ▶ Phillips Curve

Bibliography

Burchardt, F.A., et al. 1944. *The economics of full employment*. Oxford: Basil Blackwell.

Clower, R. 1965. The Keynesian counter-revolution, a theoretical appraisal. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

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

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. New York/Oxford: Oxford University Press.

Lerner, A.P. 1951. *Economics of employment*. New York: McGraw-Hill.

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

Okun, A. 1970. *The political economy of prosperity*. Washington, DC/New York: Brookings Institution/Norton.

Patinkin, D. 1956. *Money, interest, and prices*. Evanston: Row, Peterson.

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Pigou, A.C. 1933. *The theory of unemployment*. London: Macmillan.

Worswick, G.D.N. 1985. Jobs for all? *Economic Journal* 93: 1–14.

Empty Boxes

N. F. R. Crafts

'Empty Economic Boxes' is the title of a famous article in the *Economic Journal* of 1922 and a phrase which has subsequently entered the language of economics as a shorthand for 'abstract theory without practical relevance'. The paper was written by J.H. Clapham, the leading British economic historian of the interwar years and first Professor of Economic History at Cambridge University.

As an economic historian, Clapham showed a 'special predilection for hard and tangible facts' (Postan 1946, p. 57). His classic text, *An Economic History of Modern Britain* (3 vols.: 1926; 1932; 1938) is notable for its pioneering insistence on presenting facts as far as possible in detailed quantitative form. Clapham is especially remembered for his statement on the methodology of economic history subsequently eagerly embraced by 'new economic historians': Every economic historian should, however, have acquired what might be called the statistical sense, the habit of asking in relation to any institution, policy, group or movement the questions: how large? how long? how often? how representative? (1931, p. 328).

Clapham was not (nor was any of his contemporaries), however, a new economic historian, in that his work was not characterized by the use of formal economic analysis or econometrics. His approach to economic history echoes his 1922 article in finding little of practical relevance in the economic theory of his day. In effect, Clapham revealed a preference for a shift in the balance of economists' research programmes away from pure, abstract theory towards collection of better economic data.

The 1922 article was directed particularly towards Pigou's (1920) proposals for taxation of decreasing returns and subsidy of increasing returns industries. Clapham argued that 'the Laws of Returns have never been attached to specific industries; that the boxes are, in fact, empty' (1922, p. 312). In effect, Clapham criticized the lack of evidence on long-run cost curves for both firm and industry and on learning effects. He also expressed scepticism about the prospects for empirical work in this area. Pigou's reply (1922) expressed the belief that evidence would be forthcoming and that, in any case, economic thinking was useful as a method of analysis in policy questions.

Ironically, Pigou's work gave rise to a considerable body of theoretical rather than empirical literature, which left relatively little of his initial proposals intact. Notable articles by Knight (1924), Robertson (1924), Viner (1931) and Ellis and Fellner (1943) made clear, for example, the distinctions between external diseconomies and transfer payments to fixed factors and between externalities from irreversible learning effects and minimum efficient scale on a given long-run average cost curve.

Already by the end of the interwar period Pigou's hopes for empirical research were starting to be fulfilled by pioneering investigations into production functions, cost curves and learning effects, which are reviewed in Walters (1963), Johnston (1960) and Alchian (1963). It should be said that econometric work has by now achieved far more than Clapham believed possible, and the results are seriously considered in the context of antitrust policy (e.g. Cmnd. 7198, 1978), if not in taxation policy. At the same time, industrial economics textbooks still offer very substantial reservations about the precision of available estimates (Hay and Morris 1979, ch. 2).

In economic history also, there has been progress in applying Pigou's ideas as revised by the subsequent theoretical literature. By far the most interesting study is that of David (1975, ch. 2), which found econometric evidence of external economies from irreversible learning effects in the American cotton textile industry before 1825 and *also* that thereafter there was no justification for infant-industry protection.

David's paper is an excellent example of the 'new economic history', with its stress on use of theory and hypothesis testing. In fact, new economic history has generally used models based on mainstream neoclassical economics and, in general, this has undoubtedly proved fruitful – much more so than a sceptic like Clapham would have imagined. For example, discussions of the slave economy in the United States and entrepreneurial behaviour in late Victorian Britain have been substantially enriched.

Nevertheless, there is cause for concern about the one-way relationship which has developed in the past quarter-century between economics and economic history. There is a danger not so much of inevitably empty boxes as of forcing historical examples into particular boxes; that is, of operating with priors that are too tight. In particular, as McClelland (1975, p. 125) has emphasized, new economic historians should be wary of automatically believing that the marginal equivalences of the neoclassical model are tolerably achieved in all situations.

Perhaps, also, economists have more to learn from economic history than they seem presently to believe. Obviously, the past offers a much wider array of facts and institutions than the present, evidence which at present is underutilized. In addition, the study of history necessarily involves seeking to understand particular events, and exposure to the difficulties of this can give an interesting perspective on modern economic analysis. Economic historians like David would argue that the past is characterized by pervasive learning effects and technical interrelatedness so as to

produce path-dependent sequences of economic changes. If in fact the increasing-returns box is as full as initial investigations in technological history suggest it could be, then critics of orthodox theory like Kaldor (1972) will find their positions strengthened.

Moreover, in such a world the past matters in ways that neoclassical theory ignores, and the balance of research in economic history should be different; for example, less emphasis in research on the rationality of individual entrepreneurs in Victorian Britain and more on the impact of an 'early start' on subsequent economic performance.

A modern-day Clapham would still say that we do not know enough about 'increasing returns', but rather than turning his back on the concept, he would surely insist on the importance of economic history in establishing how full this box is and would now recognize that to some extent the importance of economic history depends on the answer. He would also find common cause with applied economists and econometricians like Leontief (1971) and Hendry (1980) in wishing that more resources were devoted to gathering information on economies past and present.

See Also

- ▶ Clapham, John Harold (1873–1946)
- ▶ Firm, Theory of the
- ▶ Increasing Returns to Scale

Bibliography

Alchian, A. 1963. Reliability of progress curves in airframe production. *Econometrica* 31(4): 679–693.

Clapham, J.H. 1922. Of empty economic boxes. *Economic Journal* 32: 305–314.

Clapham, J.H. 1926. *An economic history of modern Britain. Vol. 1: The early railway age*. Cambridge: Cambridge University Press.

Clapham, J.H. 1931. Economic history as a discipline. In *Encyclopaedia of the social sciences*, vol. 5, ed. E.R.A. Seligman and A. Johnson. London: Macmillan.

Clapham, J.H. 1932. *An economic history of modern Britain. Vol. 2: Free trade and steel*. Cambridge: Cambridge University Press.

Clapham, J.H. 1938. *An economic history of modern Britain. Vol. 3: Machines and national rivalries*. Cambridge: Cambridge University Press.

Command Paper no. 7198. 1978. *A review of monopolies and mergers policy*. London: HMSO.

David, P.A. 1975. *Technical choice, innovation and economic growth*. Cambridge: Cambridge University Press.

Ellis, H.S., and W. Fellner. 1943. External economies and diseconomies. *American Economic Review* 33(3): 493–511.

Hay, D.A., and D.J. Morris. 1979. *Industrial economics: Theory and evidence*. Oxford: Oxford University Press.

Hendry, D.F. 1980. Econometrics: Alchemy or science? *Economica* 47(188): 387–406.

Johnston, J. 1960. *Statistical cost analysis*. New York: McGraw-Hill.

Kaldor, N. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82: 1237–1255.

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38: 582–606.

Leontief, W. 1971. Theoretical assumptions and nonobserved facts. *American Economic Review* 61(1): 1–7.

McClelland, P.D. 1975. *Causal explanation and model building in history, economics and the new economic history*. Ithaca: Cornell University Press.

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Pigou, A.C. 1922. Empty economic boxes: A reply. *Economic Journal* 32: 458–465.

Robertson, D.H. 1924. Those empty boxes. *Economic Journal* 34: 16–30.

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3: 23–46.

Walters, A.A. 1963. Production functions and cost functions: An econometric survey. *Econometrica* 31(1): 1–66.

Encompassing

Grayham E. Mizon

Abstract

The concept of encompassing is defined and the role that it and congruence have in econometric modelling is discussed. Empirically, more than one model can appear to be congruent, but that which encompasses its rivals is dominant and will encompass all models nested within it and accurately predict the

mis-specifications of non-congruent models. These results are consistent with a general-to-specific modelling strategy being successful in practice. Alternative forms and applications of encompassing tests are discussed.

Keywords

Congruence; Encompassing; Gaussian linear regression models; Indirect inference; Models; Non-nested hypotheses; Simulation; Testing

JEL Classification

C1

Introduction and Motivation

Imaginative and productive disciplines like economics generate many new theories, partly to extend the range of phenomena that they embrace but also to improve on existing theories. New theories require rigorous evaluation to establish their worth if they are to be relevant, reliable, and robust. In addition to checking their logical consistency and relevance it is important to assess their coherence with observation. The latter usually involves the development of a model that embodies the essential characteristics of the theory and has observable implications.

The analysis presented here concentrates on the evaluation of empirical models. Numerous criteria have been proposed for assessing the coherence of an empirical model with observation. Measures of goodness of fit and selection criteria based on likelihood functions (usually degrees of freedom adjusted) are common (Schwarz 1978), and are often used both to assess coherence with observation and to select the preferred model. Probably the most comprehensive and demanding criterion for data coherence is that of congruence (Hendry 1995; Bontemps and Mizon 2003), which requires a model to be a valid reduction of whatever process actually generates the observed data – the data generation process (DGP). When $\mathbf{x}_t$ contains the full set of variables involved in an investigation, let the DGP be denoted by the joint density $D_x(\mathbf{x}_t|\mathbf{X}_{t-1}, f)$ for $\mathbf{x}_t$ conditional on its history $\mathbf{X}_{t-1}$ with parameters φ. Knowledge of the DGP endows one with omniscience and in particular the ability to derive the properties of all models involving the same variables such as $f_x(\mathbf{x}_t|\mathbf{X}_{t-1}, x)$, but, alas, for practical purposes it is unattainable. In empirical modelling, therefore, congruence means that, given the available information, the model is indistinguishable from the DGP for the chosen variables, that is, no evidence has been evinced that the model is not the DGP. Testing the latter requires that extensive, not limited, searching is done for evidence of non-congruence. This leads to the adoption of statistical tests of model mis-specification (for example, wrong functional form, heteroskedastic or serially correlated residuals) as indirect but practical tests of congruence (Hendry 1995; Mizon 1995). Since in practice a congruent model will not be the DGP, it will not necessarily be able to explain the properties of other models, and in particular those that constitute the current best knowledge and practice. Thus, a valuable part of the evaluation of a model is an assessment of whether it represents an advance on existing knowledge. 'The encompassing principle is concerned with the ability of a model to account for the behaviour of others, or less ambitiously, to explain the behaviour of relevant characteristics of other models' (Mizon 1984, p. 136). A well-known illustration in physics, discussed by Okasha (2002), for example, is provided by Newton's laws of motion and gravitation that encompassed Kepler's laws of motion and gravitation as well as Galileo's law of free-fall, and as a result the same laws explained the motion of bodies in both the terrestrial and the celestial domains. This added credence to Newton's laws, as it does for all models that encompass their rivals. It was widely believed for a long time that Newton's theory revealed the workings of nature and had the ability to explain everything in principle. However, Newton's laws have been superseded or encompassed by Einstein's relativity theory and quantum mechanics. This illustrates the fact that modelling, like discovery, is not a once-for-all event, but a continuous process of development. Progress in science, however, is achieved in many

ways, with confidence and persistence playing a role in some instances as a consequence of rejection not being accepted as final or corroboration of models that are subsequently superseded not being taken as definitive.

Background

The idea underlying the encompassing principle has a long pedigree; for example, the comparison of competing theories has been long recognized as a basic ingredient of a scientific research strategy (Nagel 1961). The implementation via a statistical contrast equally has a long history; Cox (1961, 1962) are the most significant early examples. These papers introduced statistical tests for separate families of hypotheses, and discussed several examples to illustrate their practical relevance. The tests were later developed in the literature on non-nested hypothesis testing (Pesaran 1974; Davidson and MacKinnon 1981), and encompassing (Mizon 1984). The latter paper contains a general presentation of the concept of encompassing and discussion of numerous applications, and Mizon and Richard (1986) provides a theoretical framework for encompassing, on which other theoretical papers have built extensions. Davidson et al. (1978) is one of the first attempts to develop a framework for a scientific comparison of alternative economic theories and econometric models implementing them. Different econometric models for the series of UK consumption, which rely on different economic hypotheses about consumption behaviour, were embedded in a general model and shown to imply different testable restrictions on its coefficients.

Distinguished natural scientists have expressed surprise that social scientists are able to learn anything from empirical observation when they rarely have experimental evidence. However, the encompassing principle provides precisely the analogue of the physical experiment. Experiments enable physicists and chemists to sift through alternative theories by evaluating the veracity of their implications or predictions in controlled conditions, and thus to eliminate those theories whose predictions perform badly. Congruence is the analogue of setting up controlled experimental conditions. The need to distinguish between alternative theories that each appear to be coherent with outcomes, experimental or non-experimental, leads to the search for dominant theories. For disciplines that are largely non-experimental, having a principle such as encompassing is essential for discriminating between alternative models. Typically, alternative empirical models use different information sets and possibly different functional forms, and are thus separate or non-nested. This non-nested feature enables more than one model to be congruent with respect to sample information – each can be congruent with respect to its own information set – and so it is important to assess their relative merits. Using the encompassing principle, Ericsson and Hendry (1999) analyse this issue and show that the corroboration of more than one model can imply the inadequacy of each, and Mizon (1989) provides an illustration by comparing a Keynesian and a monetarist model of inflation. Hence, congruence and encompassing are inextricably linked; in particular, encompassing comparisons of non-congruent models can be misleading. For example, general models will not always encompass simplifications of themselves even though that might seem to be an obvious characteristic of a general model, but a congruent general model will always encompass simpler models (Hendry 1995; Gouriéroux and Monfort 1995; Bontemps and Mizon 2003).

Principle

Underlying all empirical econometric analyses is an information set (collection of variables or their sigma field), and a corresponding probability space. This information set has to be sufficiently general to include all the variables thought to be relevant to the empirical implementation of theoretical models in the form of statistical models. It is also important that this information set include the variables needed for all competing models that are to be compared. When these variables are $\mathbf{x}_t$ the

DGP for the observed sample is the joint density $D_x(\mathbf{x}_t|\mathbf{X}_{t-1},f)$ at the particular parameter value $\varphi=\varphi_0$. Let a parametric statistical model of the joint distribution be $M_f=\{f_x(\mathbf{x}_t|\mathbf{X}_{t-1},x)x\in\Xi\subset R^k\}$. Let $\hat{x}$ be the maximum likelihood estimator of $\boldsymbol{\xi}$ so that $\hat{x}\xrightarrow[M_f]{P}x$ and $\hat{x}\xrightarrow[DGP]{P}x(f_0)=x_0$ which is the pseudo-true value of $\hat{x}$.

Note that the parameters of a model are not arbitrary in that M_f and its parameterization $\boldsymbol{\xi}$ are chosen to correspond to phenomena of interest such as elasticities and partial responses within the chosen probability space. For the two alternative models $M_1=\{f_1(\mathbf{x}_t|\mathbf{X}_{t-1},q_1),q_1\in\Theta_1\subset R^{p_1}\}$ and $M_2=\{f_2(\mathbf{x}_t|\mathbf{X}_{t-1},q_2),q_2\in\Theta_2\subset R^{p_2}\}$ the concept of parametric encompassing, in accordance with the approach in Mizon (1984), Mizon and Richard (1986), and Hendry and Richard (1989), can be defined as follows. M_1 encompasses M_2 (denoted $M_1\mathscr{E}M_2$) if and only if $q_{20}=\mathbf{h}_{21}(q_{10})$ when θ_{i0} is the pseudo-true value of the maximum likelihood estimator $\widehat{q}_i$ of $q_i i=1,2$, and $\mathbf{h}_{21}(q_{10})$ is the binding function given by $\widehat{q}_2\xrightarrow[M_1]{P}\mathbf{h}_{21}(q_{10})$ (Mizon and Richard 1986; Hendry and Richard 1989; Gouriéroux and Monfort 1995). Note that this definition of encompassing applies when M_1 and M_2 are non-nested as well as nested. However, Hendry and Richard (1989) showed that when M_1 and M_2 are non-nested $(M_1\mathscr{E}M_2)$ is equivalent to M_1 being a valid reduction of the minimum completing model $M_c=M_1\cup M_2^{\perp}$ (so that $M_1,M_2\subset M_c$) when $M_2{}^{\perp}$ is the model which represents all aspects of M_2 that are not contained in M_1. When this condition is satisfied, M_1 is said to parsimoniously encompass M_1 $\left(M_1\mathscr{E}_pM_c\right)$. Parsimonious encompassing is the property that a model is a valid reduction of a more general model. When a general-to-simple modelling strategy is adopted, the general unrestricted model (GUM) will have been chosen to embed the different econometric models implementing rival economic theories for the phenomenon of interest. Hence searching for the model that parsimoniously encompasses the congruent GUM is an efficient way to find congruent and encompassing models in practice. Hendry and Krolzig (2003) describe and illustrate the performance of a computer program that implements a general-to specific modelling strategy.

The comparison of Gaussian linear regression models provides a simple and convenient framework to illustrate the main ideas. Consider the two models M_1 and M_2 defined in:

$$\begin{array}{ll} M_1\backsim\mathbf{y}=\mathbf{Z}_1b+\mathbf{u}_1, & \mathbf{u}_1\backsim N\left(0,s_1^2\mathbf{I}_n\right)\\ M_2\backsim\mathbf{y}=\mathbf{Z}_2g+\mathbf{u}_2, & \mathbf{u}_2\backsim N\left(0,s_2^2\mathbf{I}_n\right)\\ M_c\backsim\mathbf{y}=\mathbf{Z}_1\mathbf{b}+\mathbf{Z}_2\mathbf{c}+e & e\backsim N\left(0,s_c^2\mathbf{I}_n\right)\end{array}\quad(1)$$

when $\mathbf{y}$ is $n\times1$, and $\mathbf{Z}_i$ is $n\times k_i(i=1,2)$ containing n observations on the independent and two sets of explanatory variables respectively with no variables in common. The explanatory variables are distributed independently of the error vectors $\mathbf{u}$, $\mathbf{v}$, and ε. When M_1, M_2 and M_c are each hypotheses about the distribution of $y|\mathbf{z}$, the models M_1 and M_2 are non-nested in that neither is a special case of the other, whereas both M_1 and M_2 are nested within M_c. A test of the hypothesis that M_1 encompasses M_2 $(M_1\mathscr{E}M_2)$ is possible using the contrast $\widehat{\psi}_\gamma=\widehat{\gamma}-\widehat{\gamma}_1=\left(\mathbf{Z}_2'\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\mathbf{Q}_1\mathbf{y}$ with $\mathbf{Q}_1=\left(\mathbf{I}_n-\mathbf{Z}_1\left(\mathbf{Z}_1'\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'\right)$ between the maximum likelihood estimator of γ, $\widehat{g}=\left(\mathbf{Z}_1'\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\mathbf{y}$, and an estimate $\widehat{g}_1\left(\mathbf{Z}_2'\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\left(\mathbf{Z}_1'\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1'\mathbf{y}$ of the pseudo-true value of $\hat{g}$ under M_1 given by $g_1=plim_{n\to\infty M_1}(\widehat{g})$. The sample complete parametric encompassing test statistic is given by $\eta_c=\widehat{\psi}_\gamma'\left(\mathbf{Z}_2'\mathbf{Z}_2\right)\left(\mathbf{Z}_2'\mathbf{Q}_1\mathbf{Z}_2\right)^{-1}\left(\mathbf{Z}_2'\mathbf{Z}_2\right)\widehat{\psi}_\gamma/k_2\widehat{\sigma}_c^2$ when $\widehat{\sigma}_c^2=\mathbf{y}'\left(\mathbf{I}_n-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\right)\mathbf{y}/(n-k_1-k_2)$ is the unbiased estimator of σ_c^2 with $\mathbf{Z}=(\mathbf{Z}_1,\mathbf{Z}_2)$. Under the complete parametric encompassing hypothesis $H_c:\psi_\gamma=\gamma-\gamma_1=0$ the statistic η_c is distributed as $F(k_2,n-k_1-k_2)$: Mizon and Richard (1986) showed that this is precisely the same statistic as that for testing the hypothesis $\mathbf{c}=0$ in (1), that is, the test statistic for $(M_1\mathscr{E}M_2)$ is exactly the same as that for $M_1\mathscr{E}_pM_c$ in this case. Variance encompassing is based on the contrast $\widehat{\psi}_{\sigma_2^2}=\widehat{\sigma}_2^2-\widehat{\sigma}_{21}^2$ between $\widehat{\sigma}_2^2$ and an estimator of $\sigma_{21}^2=\sigma_2^2+\left(\sigma_1^2/n\right)b_1'\left(\mathbf{Z}_1'\mathbf{Q}_2\mathbf{Z}_1\right)b_1$

E

the pseudo-true value $\widehat{\sigma}_2^2$ under M_1 when $\mathbf{Q}_2 = \left(\mathbf{I}_n - \mathbf{Z}_2\left(\mathbf{Z}_2'\mathbf{Z}_2\right)^{-1}\mathbf{Z}_2'\right)$. Mizon and Richard (1986) showed that the resulting variance encompassing test statistic is asymptotically equivalent to each of the one degree of freedom non-nested test statistics developed by Cox (1961, 1962), Pesaran (1974), and Davidson and MacKinnon (1981), among others. The fact that variance dominance is a necessary but not a sufficient condition for variance encompassing highlights a serious limitation of choosing models on the basis of goodness-of-fit selection criteria rather than comparing the alternative models using encompassing test statistics.

Further Developments

This analysis illustrates the fact that the choice of statistic for the encompassing contrast is very important, and may depend very much on the purpose of the analysis or the nature of the models being investigated. For example, when the GUM is not easily available or the calculation of pseudo-true values for other encompassing test statistics is difficult, comparison of the forecasting abilities provides an alternative basis for an encompassing test. Although selecting models on the basis of forecast performance can be very misleading for some purposes in a non-stationary environment with regime shifts (Hendry and Mizon 2005), the concept of forecast encompassing is a valuable method of model comparison. Forecast encompassing statistics were presented by Chong and Hendry (1990), and Ericsson (1993) and Lu and Mizon (1991) extend this analysis in several directions, including multi-step ahead forecasts from nonlinear dynamic models with estimated coefficients. Similarly, when the analytic calculation of pseudo-true values is intractable simulation methods may be used to estimate the pseudo-true values and hence compute the non-nested test statistics (Hendry and Richard 1989; Pesaran and Pesaran 1993). Gouriéroux et al. (1993) developed a comprehensive framework for such simulation known as indirect inference, which allows choice of auxiliary functions as the basis for parameter estimation. A consistent estimator of the parameters involved in the encompassing contrast can be obtained when a correction based on the simulated pseudo-true values of the testing statistics is applied. This approach has the potential to extend the application of the encompassing principle enormously. The relationship between encompassing and conditional moment or m-tests (Newey 1985) is discussed in White (1994) and Lu and Mizon (1996). The possibility that the encompassing principle be used as a generator of test statistics is discussed in Mizon and Richard (1986). Govaerts et al. (1994) consider the application of encompassing in dynamic models, and Hendry and Mizon (1993) apply it to the comparison of alternative dynamic simultaneous equations models containing integrated and cointegrated variables. A Bayesian approach to encompassing is presented in Florens et al. (1996) and, as a result of using statistical procedures rather than pseudo-true values as in Mizon and Richard (1986), argues that encompassing can be interpreted as a property of model specificity analogous to that of sufficiency for statistics. The encompassing relationship between nonparametric models is considered in Bontemps et al. (2006). Finally, Hendry et al. (2008) contains a comprehensive statement and analysis of encompassing as well as many applications of the principle.

See Also

- Artificial Regressions
- Forecasting
- Model Selection
- Models
- Testing

Bibliography

Bontemps, C., and G. Mizon. 2003. Congruence and encompassing. In *Econometrics and the philosophy of economics*, ed. B. Stigum. Princeton: Princeton University Press.

Bontemps, C., J. Florens, and J. Richard. 2006. Encompassing in regression models: parametric and non-parametric procedures. In *Progressive modelling:*

Non-nested testing and encompassing, ed. M. Marcellino and G. Mizon. Oxford: Oxford University Press.
Chong, Y., and D. Hendry. 1990. Econometric evaluation of linear macro-economic models. In *Modelling economic series*, ed. C. Granger. Oxford: Clarendon.
Cox, D. 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkeley *Symposium on Mathematical Statistics and Probability* 1, 105–23. Berkeley: University of California Press.
Cox, D. 1962. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society B* 24: 406–424.
Davidson, R., and J. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49: 781–793.
Davidson, J., D. Hendry, F. Srba, and J. Yeo. 1978. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88: 661–692.
Ericsson, N. 1993. Comment on 'On the limitations of comparing mean squared forecast errors', by M. P. Clements and D. F. Hendry. *Journal of Forecasting* 12: 644–651.
Ericsson, N., and D. Hendry. 1999. Encompassing and rational expectations: How sequential corroboration can imply refutation. *Empirical Economics* 24: 1–21.
Florens, J.-P., D. Hendry, and J.-F. Richard. 1996. Encompassing and specificity. *Econometric Theory* 12: 620–656.
Gouriéroux, C., and A. Monfort. 1995. Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11: 195–228.
Gouriéroux, C., A. Monfort, and E. Renault. 1993. Indirect inference. *Journal of Applied Econometrics* 8: 85–118.
Govaerts, B., D. Hendry, and J.-F. Richard. 1994. Encompassing in stationary linear dynamic models. *Journal of Econometrics* 63: 245–270.
Hendry, D. 1995. *Dynamic econometrics*. Oxford: Oxford University Press.
Hendry, D., and H.-M. Krolzig. 2003. New developments in automatic general-tospecific modelling. In *Econometrics and the philosophy of economics*, ed. B. Stigum. Princeton: Princeton University Press.
Hendry, D., and G. Mizon. 1993. Evaluating dynamic econometric models by encompassing the VAR. In *Models, methods and applications of econometrics*, ed. P. Phillips. Oxford: Basil Blackwell.
Hendry, D., and G. Mizon. 2005. Forecasting in the presence of structural breaks and policy regime shifts. In *Identification and inference for econometric models: Festschrift in Honor of Tom Rothenberg*, ed. D. Andrews and J. Stock. Cambridge, UK: Cambridge University Press.
Hendry, D., and J.-F. Richard. 1989. Recent developments in the theory of encompassing. In *Contributions to operations research and economics. The XXth anniversary of CORE*, ed. B. Cornet and H. Tulkens. Cambridge, MA: MIT Press.
Hendry, D.F., Marcellino, M. and Mizon, G.E., eds. 2008. Oxford bulletin of economics and statistics, Special Issue on Encompassing, 70, issue s1, 711–939.
Lu, M., and G. Mizon. 1991. Forecast encompassing and model evaluation. In *Economic structural change, analysis and forecasting*, ed. P. Hackl and A. Westlund. Berlin: Springer.
Lu, M., and G. Mizon. 1996. The encompassing principle and hypothesis testing. *Econometric Theory* 12: 845–858.
Mizon, G. 1984. The encompassing approach in econometrics. In *Econometrics and quantitative economics*, ed. D. Hendry and K. Wallis. Oxford: Blackwell.
Mizon, G. 1989. The role of econometric modelling in economic analysis. *Revista Espanola de Economia* 6: 167–191.
Mizon, G. 1995. Progressive modelling of macroeconomic time series: The LSE methodology. In *Macroeconometrics: Developments, tensions and prospects*, ed. K. Hoover. Dordrecht: Kluwer Academic Press.
Mizon, G., and J.-F. Richard. 1986. The encompassing principle and its application to non-nested hypothesis tests. *Econometrica* 54: 657–678.
Nagel, E. 1961. *The structure of science*. New York: Harcourt Brace.
Newey, W. 1985. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53: 1047–1070.
Okasha, S. 2002. *Philosophy of science: A very short introduction*. Oxford: Oxford University Press.
Pesaran, M. 1974. On the general problem of model selection. *Review of Economic Studies* 41: 153–171.
Pesaran, M., and B. Pesaran. 1993. A simulation approach to the problem of computing Cox's statistic for testing non-nested models. *Journal of Econometrics* 57: 377–392.
Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
White, H. 1994. *Estimation, inference and specification analysis*. Cambridge: Cambridge University Press.

Endogeneity and Exogeneity

John Geweke

Abstract

Endogeneity and exogeneity are properties of variables in economic or econometric models. The specification of these properties in

variables is an essential component of the process of model specification. This article considers their application in the specification of, in turn, deterministic and stochastic models.

Keywords

Cowles Commission; Endogeneity and exogeneity; Model specification; Simultaneous equations models; Statistical inference

JEL Classifications

C3

Endogeneity and exogeneity are properties of variables in economic or econometric models. The specification of these properties for respective variables is an essential component of the entire process of model specification. The words have an ambiguous meaning, for they have been applied in closely related but conceptually distinct ways, particularly in the specification of stochastic models. We consider in turn the case of deterministic and stochastic models, concentrating mainly on the latter.

A deterministic economic model typically specifies restrictions to be satisfied by a vector of variables **y**. These restrictions often incorporate a second vector of variables **x**, and the restrictions themselves may hold only if **x** itself satisfies certain restrictions. The model asserts

$$\forall \mathbf{x} \in R, \;\; \mathrm{G}(\mathbf{x}, \;\; \mathbf{y}) = 0.$$

The variables **x** are exogenous and the variables **y** are endogenous. The defining distinction between **x** and **y** is that **y** may be (and generally is) restricted by **x**, but not conversely. This distinction is an essential part of the specification of the functioning of the model, as may be seen from the trivial model,

$$\forall \mathbf{x} \in \mathbf{R}^1, \;\; x + y = 0.$$

The condition $x + y = 0$ is symmetric in x and y; the further stipulation that x is exogenous and y is endogenous specifies that in the model x restricts y and not conversely, a property that cannot be derived from $x + y = 0$. In many instances the restrictions on **y** may *determine* **y**, at least for $x \in R^* \subset R$, but the existence of a *unique* solution has no bearing on the endogeneity and exogeneity of the variables.

The formal distinction between endogeneity and exogeneity in econometric models was emphasized by the Cowles Commission in its path-breaking work on the estimation of simultaneous economic relationships. The class of models it considered is contained in the specification

$$\begin{aligned} \mathbf{B}(L)\mathbf{y}(t) + \Gamma(L)\mathbf{x}(t) &= \mathbf{u}(t); \\ \mathbf{A}(L)\mathbf{u}(t) &= \boldsymbol{\varepsilon}(t); \\ \mathrm{cov}[\boldsymbol{\varepsilon}(t), \mathbf{y}(t-s)] &= \mathbf{O}, s > 0; \\ \mathrm{cov}[\boldsymbol{\varepsilon}(t), \mathbf{x}(t-s)] &= \mathbf{O}, \text{all } s; \\ \boldsymbol{\varepsilon}(t) &\sim \mathbf{IIDN}\left(\mathbf{O}, \sum\right). \end{aligned}$$

The vectors **x**(t) and **y**(t) are observed, whereas **u**(t) and **ε**(t) are underlying disturbances not observed but affecting **y**(t). The lag operator L is defined by $L\mathbf{x}(t) = \mathbf{x}(t-1)$; the roots of $|\mathbf{B}(L)|$ and $|\mathbf{A}(L)|$ are assumed to have modulus greater than 1, a stability condition guaranteeing the non-explosive behaviour of **y** given any stable path for **x**. The Cowles Commission definition of exogeneity in this model (Koopmans and Hood 1953, pp. 117–20) as set forth in Christ (1966, p. 156) is as follows:

> An exogenous variable in a stochastic model is a variable whose value in each period is statistically independent of the values of all the random disturbances in the model in all periods.

All other variables are endogenous. In the prototypical model set forth above **x** is exogenous and **y** is endogenous.

The Cowles Commission distinction between endogeneity and exogeneity applied to a specific class of models, with linear relationships and normally distributed disturbances. The exogenous variables **x** in the prototypical model have two important but quite distinct properties. First, the model may be solved to yield an expression for y(t) in terms of current and past values of **x** and **ε**,

$$\mathbf{y}(t) = \mathbf{B}(L)^{-1}\mathbf{\Gamma}(L)\mathbf{x}(t) + \mathbf{B}(L)^{-1}\mathbf{A}(L)\varepsilon(t).$$

Given suitably restricted **x**(t) (for example, all **x** uniformly bounded, or being realizations of a stationary stochastic process with finite variance) it is natural to complete the model by specifying that it is valid for all **x** meeting the restrictions, and this is often done. The variables **x** are therefore exogenous here as **x** is exogenous in a deterministic economic model. A second, distinct property of these variables is that in estimation $\mathbf{x}(t)$ $(-\infty < t < \infty)$ may be regarded as fixed, thus extending to the environment of simultaneous equation models methods of statistical inference initially designed for experimental settings. It was generally recognized that exogeneity in the prototypical model was a sufficient but not a necessary condition to justify treating variables as fixed for purposes of inference. If **u**(t) in the model is serially independent (that is, $\mathbf{A}(L) = \mathbf{I}$) then lagged values of y may also be treated as fixed for purposes of the model; this leads to the definition of 'predetermined variables' (Christ 1966, p. 227) following Koopmans and Hood (1953, pp. 117–21):

> A variable is predetermined at time t if all its current and past values are independent of the vector of current disturbances in the model, and these disturbances are serially independent.

These two properties were not explicitly distinguished in the prototypical model (Koopmans 1950; Koopmans and Hood 1953) and tended to remain merged in the literature over the next quarter-century (for example, Christ 1966; Theil 1971; Geweke 1978). By the late 1970s there had developed a tension between the two, due to the increasing sophistication of estimation procedures in nonlinear models, treatment of rational expectations, and the explicit consideration of the respective dynamic properties of endogenous and exogenous variables (Sims 1972, 1977; Geweke 1982). Engle et al. (1983), drawing on this literature and discussions at the 1979 Warwick Summer Workshop, formalized the distinction of the two properties we have discussed. Drawing on their definitions 2.3 and 2.5 and the discussions in Sims (1977) and Geweke (1982), **x** is *model exogenous* if given $\{\mathbf{x}(t),\ t \le T\} \in \mathbf{R}(T)$ the model may restrict $\{\mathrm{y}(t),\ t \le T\}$, but given

$$\{\mathbf{x}(t),\ \ t \le T + J\} \in R(T + J)$$

there are no further restrictions on $\{\mathbf{y}(t),\ t \le \mathrm{T}\}$, for any $J > 0$. If the model in fact does restrict $\{\mathbf{y}(t),\ t \le T\}$, then **y** is model endogenous. As examples consider

Model 1:

$$y(t) = ay(t-1) + bx(t) + u(t),$$
$$x(t) = cx(t-1) + \upsilon(t);$$

Model 2:

$$y(t) = ay(t-1) + bx(t) + u(t),$$
$$x(t) = cx(t-1) + dy(t) + \upsilon(t);$$

Model 3:

$$y(t) = ay(t-1)$$
$$+b\{x(t) + E[x(t)|\, x(t-s), s > 0]\} + u(t),$$
$$x(t) = cx(t-1) + \upsilon(t).$$

In each case u(t) and v(t) are mutually and serially independent, and normally distributed. The parameters are assumed to satisfy the usual stability restrictions guaranteeing that x and y have normal distributions with finite variances. In all three models y is model endogenous, and x is model exogenous in Models 1 and 3 but not 2. For estimation the situation is different. In Model 1, treating x (t), $x(t-1)$ and y $(t-1)$ as fixed simplifies inference at no cost; $y(t-1)$ is a classic predetermined variable in the sense of Koopmans and Hood (1953) and Christ (1966). Similarly in Model 2, $x(t-1)$ and $y(t-1)$ may be regarded as fixed for purposes of inference despite the fact that x and y are both model endogenous. When Model 3 is reexpressed

$$y(t) = ay(t-1) + bx(t) + bcx(t-1) + u(t),$$
$$x(t) = cx(t-1) + \upsilon(t),$$

it is clear that $x(t)$ cannot be treated as fixed if the parameters are to be estimated efficiently since there are cross-equation restrictions involving the parameter c. Model exogeneity of a variable is thus neither a necessary nor a sufficient condition for treating that variable as fixed for purposes of inference.

The condition that a set of variables can be regarded as fixed for inference can be formalized, following Engleet al. (1983) along the lines given in Geweke (1984). Let

$$\mathbf{X} \equiv [\mathbf{x}(1), \ldots, \mathbf{x}(n)] \text{ and } \mathbf{Y} \equiv [\mathbf{y}(1), ..., \mathbf{y}(n)]$$

be matrices of n observations on the variables **x** and **y** respectively. Suppose the likelihood function $L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta})$ can be reparameterized by $\lambda = F(\boldsymbol{\Theta})$ where F is a one-to-one transformation; $\lambda' = (\lambda_1, \lambda_2)'$, $(\lambda_1, \lambda_2) \in \Lambda_1 X \Lambda_2$; and the investigator's loss function depends on parameters of interest λ_1 but not nuisance parameters λ_2. Then **x** is *weakly exogenous* if

$$L(\mathbf{X}, \mathbf{Y}|\, \lambda_1, \lambda_2) = L_1(\mathbf{Y}|\, \mathbf{X}, \lambda, \lambda_1) \cdot L_2(\mathbf{X}|\, \lambda_2),$$

and in this case *y* is *weakly endogenous.* When this condition is met the expected loss function may be expressed using only L_1 $(\mathbf{Y}|\ \mathbf{X},\ \lambda_1)$, that is, **x** may be regarded as fixed for purposes of inference.

The concepts of model exogeneity and weak exogeneity play important but distinct roles in the construction, estimation, and evaluation of econometric models. The dichotomy between variables that are model exogenous and model endogenous is a global property of a model, drawing in effect a logical distinction between the inputs of the model $\{x(t),\ t \leq T\} \in R(T)$ and the set of variables restricted by the model $\{\mathbf{y}(t),\ t \leq T\}$. Since model exogeneity stipulates that $\{\mathbf{x}(t), t \leq T + J\}$ places no more restrictions on $\{\mathbf{y}(t), t \leq T\}$ than does $\{\mathbf{x}(t), t \leq T\}$, the global property of model exogeneity is in principle testable, either in the presence or absence of other restrictions imposed by the model. When conducted in the absence of most other restrictions this test is often termed a 'causality test', and its use as a test of specification was introduced by Sims (1972). The distinction between weakly exogenous and weakly endogenous variables permits a simplification of the likelihood function that depends on the subset of the model's parameters that are of interest to the investigator. It is a logical property of the model: the same results would be obtained using $L(\mathbf{X}|\ \mathbf{Y}|\ \lambda_1,\lambda_2)$, as using $L(\mathbf{Y}|\ \mathbf{X},\ \lambda_1)$. The stipulation of weak exogeneity is therefore not, by itself, testable.

See Also

- ▶ Causality in Economics and Econometrics
- ▶ Identification
- ▶ Simultaneous Equations Models

Bibliography

Christ, C.F. 1966. *Econometric models and methods*. New York: Wiley.

Engle, R.F., D.F. Hendry, and J.-F. Richard. 1983. Exogeneity. *Econometrica* 51: 277–304.

Geweke, J. 1978. Testing the exogeneity specification in the complete dynamic simultaneous equation model. *Journal of Econometrics* 7: 163–185.

Geweke, J. 1982. Causality, exogeneity, and inference. In *Advances in econometrics*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.

Geweke, J. 1984. Inference and causality. In *Handbook of econometrics*, vol. 2, ed. Z. Griliches and M.D. Intriligator. Amsterdam: North-Holland.

Koopmans, T.C. 1950. When is an equation system complete for statistical purposes? In *Statistical inference in dynamic economic models*, ed. T.C. Koopmans. New York: Wiley.

Koopmans, T.C., and W.C. Hood. 1953. The estimation of simultaneous economic relationships. In *Studies in econometric method*, ed. W.C. Hood and T.C. Koopmans. New York: Wiley.

Sims, C.A. 1972. Money, income, and causality. *American Economic Review* 62: 540–552.

Sims, C.A. 1977. Exogeneity and causal ordering in macroeconomic models. In *New methods in business cycle research*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Theil, H. 1971. *Principles of econometrics*. New York: Wiley.

Endogenous and Exogenous Money

Meghnad Desai

The issue of endogeneity or exogeneity of money is one that runs through the history of monetary theory, with prominent authors appearing to hold views on either side. Narrowly put, those who plug for the exogeneity view take one or all among the cluster of variables – price level, interest rate or real output – as being determined by movements in the stock of money. Those who

hold the endogeneity view consider that the stock of money in circulation is determined by one or all of the variables mentioned above. This narrow definition begs several questions. The variables price level (P), interest rate (R), real output (Y) and money stock (M) are all at the macroeconomic level, i.e. in the context of a one-good economy. Some part of the continuing debate can be traced to the view held by various participants in the controversy about whether such a high level of aggregation is appropriate, e.g. is there a rate of interest? Another part of the debate refers to the choice of money stock variable. Is it commodity money (gold), fiat (paper) money, bank deposits or a larger measure of liquidity that is to stand for *the* money stock? The problem can be dealt with even at a one-good level either in the context of a closed economy or an open economy and either in an equilibrium or a disequilibrium context, static or dynamic, short run or long run. The basic issue is about the direction of causality-money to other variables or other variables to money. But as our understanding of the underlying statistical theory concerning causality and exogeneity has advanced in recent years, it must also be added that participants in the controversy conflate the exogeneity of a variable (especially of money) with its *controllability* by policy. Strictly speaking one can have exogeneity without any presumption that the variable can be manipulated by policy, for example rainfall. Also once posed in a dynamic context, we should distinguish between weak exogeneity, which allows for feedback from the endogenous to the exogenous variables over time, and strong exogeneity, which does not allow such a feedback (Hendry et al. 1983). Endogeneity or exogeneity are notions that only make sense in the context of a model. Frequently in the past, there has been a failure to specify such a model, which has then allowed the controversy to continue.

Some Definitions

To simplify matters, at the risk of putting off readers, let us begin by specifying a small model within whose context endogeneity and exogeneity can be defined. This macroeconomic model will consist of four variables P, Y, R and M whose exogenous/endogenous status is at debate. We subdivide them into the three non-monetary variables P, Y, R labelled X and money M. There are of course other truly exogneous variables – tastes, technology, international variables – which we label Z. Now we observe that the variables X and M are correlated, i.e. jointly distributed conditional upon the set of variables Z. The question of endogeneity or exogeneity of money is as to whether the correlation between X and M can be written in terms of X being a function of M and Z, or M being a function of X and Z. In econometric terms, can we partition the joint distribution of X and M into a *conditional* distribution of X on M, Z and a marginal distribution of M on Z (the exogenous money case) or a conditional distribution of M on X and Z and a marginal distribution of X on Z. Thus when we say money is exogenous it is exogenous with respect to X variables but it could still be determined by Z variables; symmetrically for the X variables being exogenous. If M is influenced by the past values of X as well as by Z though not by the current values of X, then M is said to be weakly exogenous. Thus M may be controlled by monetary authorities but they may be reacting to past behaviour of X variables. Then M is determined by a reaction function and is only weakly exogenous. The same definition of weak exogeneity extends to the Z variables. Thus even international variables, such as capital inflow, may be determined by past values of X variables in which case they are weakly exogenous (for further detail, see Desai 1981). The best way to consider the issue of exogeneity of money is to specify the type of money economy envisaged – commodity money, paper money, credit money and look at the variables likely to influence the supply of money and its relation with other variables.

Commodity Money

Historically the argument about exogeneity is constructed around the Quantity Theory of Money, which stated that the amount of money in circulation at any time determined the volume

of trade and if the amount went on increasing it would lead sooner or later to an increase in price. In the context of commodity money, the proposition concerned attempts by coining authorities to debase coinage by clipping or alloying it with inferior metal. These were ways in which the amount of money could be altered by policy manipulation and then exogenously act upon prices. But in a commodity money regime, the stock of money could also be altered by influx of precious metal through gold discoveries and greater influx. These were exogenous variations not susceptible to policy manipulation but presumed an open economy. The first statement of the quantity theory of money by David Hume starts with an illustration of an influx of gold from outside and traces its effects first on real economic activity and eventually on prices. In Hume's quantity theory, money is exogenous but not subject to policy manipulation. The opposite view (argued by James Steuart for instance) was that it was the volume of activity that elicited the matching supply of money. This could be done partly by dis-hoarding on the parts of those who now expected a better yield on their stock. It could also be altered if banks were willing to 'accommodate' a larger volume of bills (see Desai 1981). Dis-hoarding implies that a portion of the money supply in *circulation* is endogenously determined in a commodity money economy. It could be argued that even the influx of gold could have been caused by the discrepancy between the domestic and the world gold price, which in the 18th century before a world gold market existed could be substantial. In the latter case money would be weakly exogenous as long as there were lags between the appearance of discrepancy and the inflow of gold.

Inside Money

Once however one introduces banks into the scheme of things, the issue of exogeneity becomes complex. Till very recently we have lacked a theory of banking behaviour of any degree of sophistication, although in terms of institutional description we have much knowledge. If banks are willing to 'accommodate' a greater volume of trade, this can only be because they find it profitable to do so. This increased profitability may be actual or perceived but it must be a result of an increase in differential between the interest (discount) rate borrowers are willing to pay and the rate at which banks can acquire liquidity. Banks can then choose to expand the ratio of credit to the cash base and sustain a higher volume. Banks create inside money and inside money can only be regarded as endogenous. But the extent to which a single bank can create money will depend on the behaviour of the *banking system*. The banking system can by the cloakroom mechanism choose any ratio of credit to cash base. It is conceivable though not likely that in such a system of inside money, banks could arbitrarily, i.e. exogenously, increase money supply. They must however base such an action on considerations of expected profitability. We can envisage a situation in which banks guided by 'false' expectations can sustain a credit boom by a bootstraps mechanism. This is the way in which a Wicksellian cumulative process could sustain itself. An arbitrary, exogenous increase in inside money by the banking system though possible is not very likely. It runs into the problems caused by the leakage of cash either internally (finite limits to the velocity of circulation of cash) or abroad. It was the international leakage that was normally regarded as the most likely constraint since it caused outflow of gold – the International Gold Standard which provided the context for 19th-century theories in this imposed exogenous constraints on money supply by imposing a uniform gold price in all countries. In such a case, money is exogenous and not subject to policy manipulation. In as much as gold movements are triggered by internal variables, it is weakly exogenous.

Outside Fiat Money

It is the case of fiat money printed as the state's liability, i.e. as outside money, that provides the best illustration of exogenous money not subject to any constraint. In a world where only paper

currency was used and it was printed by the monetary authorities, the stock of money could be exogenously determined. This would be additionally so even if there was inside money as long as the monetary authorities could insist that banks obeyed a strict cash to deposit ratio and there were no substitutes for cash available beyond the control of the monetary authorities. It is this view of money that most closely corresponds to Keynes's assumptions in the *General Theory* and it is also in the monetarist theory of Milton Friedman. The banking system is a passive agent in this view and given the cash base is always fully loaned up. Thus given the amount of high powered money in the system providable only by the monetary authorities, the supply of money is determined. Even if the stock of money were exogenous, its impact on the non-monetary variables X can be variable. This is because the velocity of circulation which translates the stock of money into money in circulation need not be constant but variable. If the velocity of circulation were not only a variable but also a function of the X variables, then although the monetary authorities can determine the stock of money the influence of money on real variables is not as predicted by the Quantity Theory. Thus it is not the exogeneity of money issue that divides monetarists and Keynesians but the determinants of the velocity of circulation. For the monetarist, the velocity of circulation ($M/P \cdot Y$) has to be independent of P, Y, R and M. For Keynesians, the demand for money depends on the rate of interest crucially and the interest elasticity of demand for money is a variable tending to infinity in a liquidity trap.

Modern Credit Economy

In a world with inside and outside money with a sophisticated banking system as well as a non-banking financial sector, the question of exogeneity is the most complex. In the previous case of outside fiat money we assumed, that the cash ratio was fixed and adhered to by banks. It is when the banks' reserve base contains government debt instruments – treasury bills, bonds, etc. – that the profit-maximizing behaviour of the banks renders a greater part of the money stock endogenous. Thus while the narrow money base – currency in circulation and in central bank reserves – can be regulated by the monetary authority, the connection between money base and total liquidity in the economy becomes highly variable. Banks will expand their loan portfolio as long as the cost of replenishing their liquidity does not exceed the interest rate they can earn on loans. The relation between broad money (M_3) and narrow money (M_0) becomes a function of the funding policy concerning the budget deficit and the structure of interest rates. Thus the stock of narrow money can be exogenous and policy determined. But the stock of broad money is endogenous. A crucial recent element has been the financial revolution of the last decade (De Cecco 1987). A variety of financial instruments – credit cards, charge cards, money market funds, interest-bearing demand deposits, electronic cash transfer – has made the ratio of cash to volume of financial transactions variable though with a steep downward trend. It has also increased the number of money substitutes and made the cost of liquidity lower. The non-banking financial system thus can create liquidity by 'accommodating' a larger volume of business, advancing trade credit, allowing consumer debt to increase etc. The velocity of circulation of cash increases very sharply in such a world and liquidity, a broader concept than even broad money, becomes endogenous. Here again profitability of liquidity creation becomes the determining variable. But the financial revolution has also integrated world financial markets and economies are increasingly open. Thus capital flows are rapid and respond to minute discrepancies in the covered interest parity. In such a world money is at best weakly exogenous but more usually endogenous. The issue of exogeneity or endogeneity of money thus crucially depends on the type of money economy that one is considering – commodity money, paper money, credit (mobile) money. It also depends on the sophistication of the banking and financial system within which such money is issued. Debates over the last two hundred years have used the word money to cover a variety of situations. It has also not been clarified whether the issue is exogeneity

of money or its controllability and whether it is merely the stock of money or its velocity as well which is being considered. Once these issues have been clarified, the notion of exogeneity needs to be defined in the modern econometric fashion, relative to a model in order to decide whether money can be exogenous. It seems likely that the narrower the definition of money stock, the more likely is it to fulfil the requirement of (weak) exogeneity. Such exogeneity is necessary but not sufficient to demonstrate that money determines the price level or the real economy.

See Also

- ▶ Capital, Credit and Money Markets
- ▶ High-Powered Money and the Monetary Base
- ▶ Monetary Base
- ▶ Money Supply
- ▶ Quantity Theory of Money

Bibliography

De Cecco, M. 1987. *Changing money: Financial innovations in developed countries*. Oxford: Blackwell.

Desai, M. 1981. *Testing monetarism*. London: Frances Pinter.

Hendry, D., R. Engle, and J.L. Richard. 1983. Exogeneity. *Econometrica* 51(2): 227–304.

Endogenous Growth Theory

Peter Howitt

Abstract

Endogenous growth theory explains long-run growth as emanating from economic activities that create new technological knowledge. This article sketches the outlines of the theory, especially the 'Schumpeterian' variety, and briefly describes how the theory has evolved in response to empirical discoveries.

Keywords

Aggregate production function; Capital accumulation; Competition; Creative destruction; Economic growth; Endogenous growth; Human capital; Innovations; Intellectual capital; Intermediate products; Intertemporal utility maximization; Law of large numbers; Marginal product of capital; Neoclassical growth theory; Physical capital; Product variety; Productivity growth; Research and development; Saving rate; Schumpeterian growth; Steady state; Technological progress; Technology; Technology frontier; Total Factor productivity; Transfer of technology

JEL Classifications

O4

Endogenous growth is long-run economic growth at a rate determined by forces that are internal to the economic system, particularly those forces governing the opportunities and incentives to create technological knowledge.

In the long run the rate of economic growth, as measured by the growth rate of output per person, depends on the growth rate of total factor productivity (TFP), which is determined in turn by the rate of technological progress. The neoclassical growth theory of Solow (1956) and Swan (1956) assumes the rate of technological progress to be determined by a scientific process that is separate from, and independent of, economic forces. Neoclassical theory thus implies that economists can take the long-run growth rate as given exogenously from outside the economic system.

Endogenous growth theory challenges this neoclassical view by proposing channels through which the rate of technological progress, and hence the long-run rate of economic growth, can be influenced by economic factors. It starts from the observation that technological progress takes place through innovations, in the form of new products, processes and markets, many of which are the result of economic activities. For example, because firms learn from experience how to produce more efficiently, a higher pace of economic

activity can raise the pace of process innovation by giving firms more production experience. Also, because many innovations result from R&D expenditures undertaken by profit-seeking firms, economic policies with respect to trade, competition, education, taxes and intellectual property can influence the rate of innovation by affecting the private costs and benefits of doing R&D.

AK Theory

The first version of endogenous growth theory was AK theory, which did not make an explicit distinction between capital accumulation and technological progress. In effect it lumped together the physical and human capital whose accumulation is studied by neoclassical theory with the intellectual capital that is accumulated when innovations occur. An early version of AK theory was produced by Frankel (1962), who argued that the aggregate production function can exhibit a constant or even increasing marginal product of capital. This is because, when firms accumulate more capital, some of that increased capital will be the intellectual capital that creates technological progress, and this technological progress will offset the tendency for the marginal product of capital to diminish.

In the special case where the marginal product of capital is exactly constant, aggregate output Y is proportional to the aggregate stock of capital K:

$$Y = AK \tag{1}$$

where A is a positive constant. Hence the term 'AK theory'.

According to AK theory, an economy's long-run growth rate depends on its saving rate. For example, if a fixed fraction s of output is saved and there is a fixed rate of depreciation δ, the rate of aggregate net investment is:

$$\frac{dK}{dt} = sY - \delta K$$

which along with (1) implies that the growth rate is given by:

$$g \equiv \frac{1}{Y}\frac{dY}{dt} = \frac{1}{K}\frac{dK}{dt} = sA - \delta.$$

Hence an increase in the saving rate s will lead to a permanently higher growth rate.

Romer (1986) produced a similar analysis with a more general production structure, under the assumption that saving is generated by intertemporal utility maximization instead of the fixed saving rate of Frankel. Lucas (1988) also produced a similar analysis focusing on human capital rather than physical capital; following Uzawa (1965) he explicitly assumed that human capital and technological knowledge were one and the same.

Innovation-Based Theory

AK theory was followed by a second wave of endogenous growth theory, generally known as 'innovation-based' growth theory, which recognizes that intellectual capital, the source of technological progress, is distinct from physical and human capital. Physical and human capital are accumulated through saving and schooling, but intellectual capital grows through innovation.

One version of innovation-based theory was initiated by Romer (1990), who assumed that aggregate productivity is an increasing function of the degree of product variety. In this theory, innovation causes productivity growth by creating new, but not necessarily improved, varieties of products. It makes use of the Dixit–Stiglitz–Ethier production function, in which final output is produced by labour and a continuum of intermediate products:

$$Y = L^{1-\alpha}\int_0^A x(i)^{\alpha}\,di, 0 < \alpha < 1 \tag{2}$$

where L is the aggregate supply of labour (assumed to be constant), $x(i)$ is the flow input of intermediate product i, and A is the measure of different intermediate products that are available for use. Intuitively, an increase in product variety, as measured by A, raises productivity by allowing society to spread its intermediate production more thinly across a larger number of activities, each of

which is subject to diminishing returns and hence exhibits a higher average product when operated at a lower intensity.

The other version of innovation-based growth theory is the 'Schumpeterian' theory developed by Aghion and Howitt (1992) and Grossman and Helpman (1991). (Early models were produced by Segerstrom et al. 1990 and Corriveau 1991). Schumpeterian theory focuses on quality-improving innovations that render old products obsolete, through the process that Schumpeter (1942) called 'creative destruction.'

In Schumpeterian theory aggregate output is again produced by a continuum of intermediate products, this time according to:

$$Y = L^{1-\alpha} \int_0^1 A(i)^{1-\alpha} x(i)^{\alpha} di, \tag{3}$$

where now there is a fixed measure of product variety, normalized to unity, and each intermediate product i has a separate productivity parameter $A(i)$. Each sector is monopolized and produces its intermediate product with a constant marginal cost of unity. The monopolist in sector i faces a demand curve given by the marginal product: $\alpha \cdot (A(i)L/x(i))^{1-\alpha}$ of that intermediate input in the final sector. Equating marginal revenue (α time this marginal product) to the marginal cost of unity yields the monopolist's profit-maximizing intermediate output:

$$x(i) = \xi L A(i)$$

where $\xi = \alpha^{2/(1-\alpha)}$. Using this to substitute for each $x(i)$ in the production function (3) yields the aggregate production function:

$$Y = \theta A L \tag{4}$$

where $\theta = \xi^{\alpha}$, and where A is the average productivity parameter:

$$A \equiv \int_0^1 A(i)\, di.$$

Innovations in Schumpeterian theory create improved versions of old products. An innovation in sector i consists of a new version whose productivity parameter A exceeds that of the previous version by the fixed factor $\gamma > 1$. Suppose that the probability of an innovation arriving in sector i over any short interval of length dt is $\mu \cdot dt$. Then the growth rate of $A(i)$ is

$$\frac{dA(i)}{A(i)} \cdot \frac{1}{dt} = \left\{ \begin{array}{ll} (\gamma - 1) \cdot \dfrac{1}{dt} & \text{with probability } \mu \cdot dt \\ 0 & \text{with probability } 1 - \mu \cdot dt \end{array} \right\}.$$

Therefore the expected growth rate of $A(i)$ is:

$$E(g) = \mu(\gamma - 1). \tag{5}$$

The flow probability μ of an innovation in any sector is proportional to the current flow of productivity-adjusted R&D expenditures:

$$\mu = \lambda R/A \tag{6}$$

where R is the amount of final output spent on R&D, and where the division by A takes into account the force of increasing complexity. That is, as technology advances it becomes more complex, and hence society must make an ever-increasing expenditure on research and development just to keep innovating at the same rate as before.

It follows from (4) that the growth rate g of aggregate output is the growth rate of the average productivity parameter A. The law of large numbers guarantees that g equals the expected growth rate (5) of each individual productivity parameter. From this and (6) we have:

$$g = (\gamma - 1)\lambda R/A.$$

From this and (4) it follows that the growth rate depends on the fraction of GDP spent on research and development, $n = R/Y$, according to:

$$g = (\gamma - 1)\lambda \theta L n. \tag{7}$$

Thus, innovation-based theory implies that the way to grow rapidly is not to save a large fraction

of output but to devote a large fraction of output to research and development. The theory is explicit about how R&D activities are influenced by various policies, who gains from technological progress, who loses, how the gains and losses depend on social arrangements, and how such arrangements affect society's willingness and ability to create and cope with technological change, the ultimate source of economic growth.

Empirical Challenges

Endogenous growth theory has been challenged on empirical grounds, but its proponents have replied with modifications of the theory that make it consistent with the critics' evidence. For example, Mankiw et al. (1992), Barro and Sala-i-Martin (1992) and Evans (1996) showed, using data from the second half of the 20th century, that most countries seem to be converging to roughly similar long-run growth rates, whereas endogenous growth theory seems to imply that, because many countries have different policies and institutions, they should have different long-run growth rates. But the Schumpeterian model of Howitt (2000), which incorporates the force of technology transfer, whereby the productivity of R&D in one country is enhanced by innovations in other countries, implies that all countries that perform R&D at a positive level should converge to parallel long-run growth paths.

The key to this convergence result is what Gerschenkron (1952) called the 'advantage of backwardness'; that is, the further a country falls behind the technology frontier, the larger is the average size of innovations, because the larger is the gap between the frontier ideas incorporated in the country's innovations and the ideas incorporated in the old technologies being replaced by innovations. This increase in the size of innovations keeps raising the laggard country's growth rate until the gap separating it from the frontier finally stabilizes.

Likewise, Jones (1995) has argued that the evidence of the United States and other OECD countries since 1950 refutes the 'scale effect' of Schumpeterian endogenous growth theory. That is, according to the growth Eq. 7 an increase in the size of population should raise long-run growth by increasing the size of the workforce L, thus providing a larger market for a successful innovator and inducing a higher rate of innovation. But in fact productivity growth has remained stationary during a period when population, and in particular the number of people engaged in R&D, has risen dramatically. The models of Dinopoulos and Thompson (1998), Peretto (1998) and Howitt (1999) counter this criticism by incorporating Young's (1998) insight that, as an economy grows, proliferation of product varieties reduces the effectiveness of R&D aimed at quality improvement by causing it to be spread more thinly over a larger number of different sectors. When modified this way the theory is consistent with the observed coexistence of stationary TFP growth and rising population, because in a steady state the growth-enhancing scale effect is just offset by the growth-reducing effect of product proliferation.

As a final example, early versions of innovation-based growth theory implied, counter to much evidence, that growth would be adversely affected by stronger competition laws, which by reducing the profits that imperfectly competitive firms can earn ought to reduce the incentive to innovate. However, Aghion and Howitt (1998, ch. 7) describe a variety of channels through which competition might in fact spur economic growth. One such channel is provided by the work of Aghion et al. (2001), who show that, although an increase in the intensity of competition will tend to reduce the absolute level of profits realized by a successful innovator, it will nevertheless tend to reduce the profits of an unsuccessful innovator by even more. In this variant of Schumpeterian theory, more intense competition can have a positive effect on the rate of innovation because firms will want to escape the competition that they would face if they lost whatever technological advantage they have over their rivals.

Much more work needs to be done before we can claim to have a reliable explanation for why economic growth is faster in some countries and in some time periods than in others. But the fact that much of the cross-country variation in growth rates

is attributable to differences in productivity growth rather than differences in rates of capital accumulation suggests that endogenous growth theory, which aims to provide an economic explanation of these differences in productivity growth, will continue to attract economists' attention for years to come.

See Also

▶ Schumpeterian Growth and Growth Policy Design

Bibliography

Aghion, P., and P. Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60: 323–351.
Aghion, P., and P. Howitt. 1998. *Endogenous growth theory*. Cambridge, MA: MIT Press.
Aghion, P., C. Harris, P. Howitt, and J. Vickers. 2001. Competition, imitation and growth with step-by-step innovation. *Review of Economic Studies* 68: 467–492.
Barro, R.J., and X. Sala-i-Martin. 1992. Convergence. *Journal of Political Economy* 100: 223–251.
Corriveau, L. 1991. Entrepreneurs, growth, and cycles. Doctoral dissertation, University of Western Ontario.
Dinopoulos, E., and P. Thompson. 1998. Schumpeterian growth without scale effects. *Journal of Economic Growth* 3: 313–335.
Evans, P. 1996. Using cross-country variances to evaluate growth theories. *Journal of Economic Dynamics and Control* 20: 1027–1049.
Frankel, M. 1962. The production function in allocation and growth: A synthesis. *American Economic Review* 52: 995–1022.
Gerschenkron, A. 1952. Economic backwardness in historical perspective. In *The progress of underdeveloped areas*, ed. B.F. Hoselitz. Chicago: University of Chicago Press.
Grossman, G.M., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.
Howitt, P. 1999. Steady endogenous growth with population and R&D inputs growing. *Journal of Political Economy* 107: 715–730.
Howitt, P. 2000. Endogenous growth and cross-country income differences. *American Economic Review* 90: 829–846.
Jones, C.I. 1995. R&D-based models of economic growth. *Journal of Political Economy* 103: 759–784.
Lucas, R.E. Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.
Mankiw, N.G., D. Romer, and D.N. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–437.
Peretto, P.F. 1998. Technological change and population growth. *Journal of Economic Growth* 3: 283–311.
Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037.
Romer, P.M. 1990. Endogenous technological change. *Journal of Political Economy* 98: S71–S102.
Schumpeter, J.A. 1942. *Capitalism, socialism and democracy*. New York: Harper.
Segerstrom, P.S., T.C.A. Anant, and E. Dinopoulos. 1990. A Schumpeterian model of the product life cycle. *American Economic Review* 80: 1077–1091.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Uzawa, H. 1965. Optimal technical change in an aggregative model of economic growth. *International Economic Review* 6: 18–31.
Young, A. 1998. Growth without scale effects. *Journal of Political Economy* 106: 41–63.

Endogenous Market Incompleteness

Christopher Sleet

Abstract

Endogenously incomplete models derive restrictions on asset trading from primitive constraints on the enforcement and monitoring technologies available to societies. They have been applied to a wide variety of macroeconomic problems. This article reviews some of these applications and the models that underpin them.

Keywords

Asset pricing models; Autarky; Bilateral insurance games; Commitment; Default; Dynamic moral hazard models; Endogenously incomplete markets; Equity premium; Euler equations; Incomplete markets; Information revelation; Limited enforcement models; Risk sharing; Taxation of assets

JEL Classifications

D4; D10

An asset trading arrangement is incomplete if it is too restrictive to ensure a fully Pareto-optimal allocation of risk. Endogenously incomplete models derive such trading arrangements from primitive frictions. They are to be contrasted with models that *assume* a particular incomplete asset markets structure.

Recent contributions to the endogenous incompleteness literature have emphasized imperfections in the enforcement and monitoring technologies available to societies. They derive endogenous market structures, sometimes supplemented with a tax system, as decentralizations of planning problems in which the planner faces one or both of these imperfections. These market-tax structures ensure that agents are provided with incentives to honour promises that cannot be costlessly enforced or that are contingent on states that cannot be costlessly observed. By construction they admit equilibria that are constrained efficient.

Models with endogenous incompleteness have received a variety of applications in macroeconomics. They have been used to enhance understanding of risk sharing, asset pricing and business cycles; on the normative side they have been applied to analyses of optimal fiscal policy. Here I review some of these applications and the models that underpin them.

Limited Enforcement

The canonical example of a limited enforcement model is the bilateral insurance game of Kocherlakota (1996). In this game, two risk-averse agents are endowed with random and imperfectly correlated income processes. Neither agent can be compelled to deliver resources to the other, even if they have promised to do so in the past.

Equilibrium allocations in this setting can be implemented with strategies that revert to autarky following an agent defection. Agents with high-income shocks can be induced to share some of their resources by the threat of such reversion and, when this is insufficient, by promises of extra resources in the future. Such promises introduce additional dynamics into optimal equilibrium allocations; shocks that cannot be smoothed over states are smoothed over time instead, ensuring that individual consumption is persistent even when aggregate consumption is not.

Constrained-efficient allocations in limited enforcement economies can be decentralized using a complete set of Arrow security markets coupled with endogenous debt limits (see Alvarez and Jermann 2000). Intuitively, agents can borrow only up to the amount that they are willing to pay back in the future given that the penalty for default is consignment to autarky. Thus, the limited enforcement friction provides a micro-foundation for the often-made assumption of a debt limit tighter than that implied by an agent's intertemporal budget constraint. In the limited enforcement case, however, the debt limit is state-contingent; it depends upon the value of autarky to the agent. Since this value is a function of individual and aggregate shocks, the parameters of the shock process and, in richer models, the agent's opportunities for self- or public insurance after exclusion from markets, so too is the debt limit.

When agents' endogenous debt limits periodically bind, risk sharing is disrupted; individual consumption, conditional on the aggregate state, is positively correlated with current and past individual income. Qualitatively, such departures from full risk-sharing cohere well with evidence on individual consumption. In Alvarez and Jermann's (2001) quantitative analysis of a calibrated limited enforcement model, the endogenous debt limits bind fairly often and permit relatively little risk sharing. This is consistent with evidence on the sharing of low- frequency risks. Alvarez and Jermann's analysis also has implications for asset pricing. They obtain a volatile asset pricing kernel and risk premia that are large and time varying. These implications are consistent with asset pricing data, but contrast with those of the benchmark representative agent asset pricing model.

Cross-country consumption data also exhibit apparent departures from full risk sharing. Standard models (with complete markets) imply co-movements in consumption that exceed those in output, yet the data suggests the reverse. Kehoe

and Perri (2002) show that a limited enforcement model augmented with production and physical capital accumulation can go some way to explaining this anomaly.

Recent papers have considered alternative penalties for default including the confiscation of an endogenously valued collateral asset (see, Lustig 2005) or the payment of a fixed default cost (Cooley et al. 2004). These contributions illustrate the scope of limited enforcement models: Lustig explores the implications of endogenously valued collateral for asset pricing and obtains a large and time-varying price of risk; Cooley, Marimon and Quadrini examine the role of limited enforcement frictions in propagating business cycle shocks. Cordoba (2005) and Arpad and Cárceles-Poveda (2005), however, sound cautionary notes. They provide calibrated models in which the introduction of collateral relaxes endogenous debt limits so much that agents can fully diversify risk.

Private Information

An alternative line of research has analysed environments in which risk-averse agents privately observe shocks to their endowments, tastes or productivity (see, for example, Atkeson and Lucas 1992). In this setting, agents must be provided with incentives to reveal information. The socially efficient provision of incentives requires the conditioning of current consumption on an agent's history of shock reports. Intuitively, agents are rewarded for reporting a low current need for resources with the promise of more consumption in the future. Thus, intertemporal consumption smoothing is enhanced and interstate smoothing disrupted.

Albanesi and Sleet (2006) and Kocherlakota (2005) show that optimal information-constrained allocations can be implemented with a mixture of non-contingent debt markets and taxes. Thus, these authors derive joint restrictions on the market structure *and* the tax system from primitive informational frictions. Central to their analyses is an 'inverted Euler equation'. If $\{c_t^*\}_{t=0}^{\infty}$ denotes the optimal consumption allocation, this equation is given by:

$$\frac{1}{u'\left(c_t^*(z^t,\theta^t)\right)} = \beta\lambda_{t+1}(z^{t+1})E_t\left[\frac{1}{u'\left(c_{t+1}^*\right)}|z^{t+1},\theta^t\right]. \tag{1}$$

Here θ^t denotes an agent's period t history of privately observed shocks, z and z^{t+1} denote t and $t+1$ histories of observable aggregate shocks, β is the agent's discount factor and u' her marginal utility of consumption. λ_{t+1} is a social stochastic discount factor (SSDF) that 'prices' resources delivered after each history z^{t+1}. Golosov et al. (2003) show that such equations hold in a large class of dynamic moral hazard models. They imply a wedge between an agent's conditional expected intertemporal marginal rate of substitution (IMRS) and the SSDF. This wedge provides a rationale for asset taxation; intuitively, agents must be discouraged from saving at date t since greater wealth at $t+1$ undermines incentives at that date. However, the implications for asset taxation are subtle. The optimal allocation cannot be implemented with an asset tax that merely 'matches the wedge' and equates the conditional expectation of an agent's IMRS to the SSDF. Instead, marginal asset taxes at $t+1$ are used to generate a positive covariance between the after-tax asset return and the agent's consumption that deters savings. In some cases, the expected asset tax is zero and the wedge is entirely generated by this covariance effect.

Positive analyses of dynamic moral hazard are relatively scarce. Green and Oh (1991) contrast the empirical implications of various incomplete market models, including those with moral hazard. Kocherlakota and Pistaferri (2005) identify λ_{t+1} with the market discount factor, assume that utility has the constant relative risk aversion property and use (1) to derive expressions for λ_{t+1} in terms of cross-sectional moments of the consumption distribution. They then investigate the implications of this dynamic moral hazard model for asset pricing and, in particular, the equity premium and risk-free rate. They find that plausible

values of the coefficient of relative risk aversion set the equity premium pricing error to zero.

In all of the dynamic moral hazard models described so far, the consumption of agents is observable. An alternative assumption is that agents can undertake asset trades that are hidden from society. Agents must now be given incentives to reveal information *and* save an appropriate amount. This places additional constraints on risk sharing. When agents can control their publicly observable histories and can save at the prices implied by an exogenously given sequence of SSDFs, these constraints are severe. In this case, the optimal allocation is identical to that in an economy with riskless debt (see Cole and Kocherlakota 2001). This result is important as it provides a micro-foundation for models that exogenously restrict agents to the trading of such debt.

Government Incentive Problems

Governments or mechanism designers may also have difficulty keeping their promises. There is a long tradition of considering commitment problems in Ramsey models. In these, a socially benevolent government typically has access to a restricted set of linear tax mechanisms and an asset market in which it can trade claims to resources. *Ex ante* optimal policy entails implicit promises over future allocations and, in particular, the expected value of the government's future stream of primary surpluses that it is rarely in the government's interests to keep. For example, if the government can default on its debt it will, since in this way it can avoid the distortionary taxes necessary for debt repayment. As in the limited-enforcement models described above, reversion to autarky after a default can sustain some equilibrium borrowing by the government, though typically it implies a tight endogenous debt limit (Chari and Kehoe 1993). Sleet (2004) and Sleet and Yeltekin (2006a) consider models in which the government's true spending needs are not publicly observable. Although the government has access to a complete set of contingent claims markets, in equilibrium it is required to adopt a debt-trading policy consistent with truthful revelation of its spending needs. This limits its ability to buy claims against high spending-needs states and sell them against low spending-needs ones. The outcome is enhanced intertemporal, as opposed to inter-state, smoothing of taxes.

The optimal allocations and market-tax implementations implied by dynamic moral hazard models also involve promises from a planner (or government) to an agent. These allocations often entail the absorption of almost all agents by a minimal utility immiserating state; they thus place strong demands on the planner's ability to commit. Sleet and Yeltekin (2006b) remove this ability. They show that optimal allocations without planner commitment solve the problems of *committed* planners who discount the future *less heavily* than agents. Coupling this result with the work of Farhi and Werning (2005), who directly assume a planner discount factor in excess of the agents, suggests that constrained optimal allocations can be implemented with non-contingent debt, an income tax and a progressive estate tax. Analysis of dynamic moral hazard models without societal commitment is, however, still in its infancy and much remains to be done.

See Also

- ▶ Default and Enforcement Constraints
- ▶ Optimal Fiscal and Monetary Policy (Without Commitment)
- ▶ Social Insurance

Bibliography

Albanesi, S., and C. Sleet. 2006. Dynamic optimal taxation with private information. *Review of Economic Studies* 73: 1–30.

Alvarez, F., and U. Jermann. 2000. Efficiency, equilibrium and asset pricing with risk of default. *Econometrica* 68: 775–797.

Alvarez, F., and U. Jermann. 2001. Quantitative asset pricing implications of endogenous solvency constraints. *Review of Financial Studies* 14: 1117–1151.

Arpad, A., and E. Cárceles-Poveda. 2005. *Endogenous trading constraints with incomplete markets*, Working paper. University of Rochester.

Atkeson, A., and R. Lucas. 1992. On efficient distribution with private information. *Review of Economic Studies* 59: 427–453.
Chari, V., and P. Kehoe. 1993. Sustainable plans and debt. *Journal of Economic Theory* 60: 175–195.
Cole, H., and N. Kocherlakota. 2001. Efficient allocations with hidden income and storage. *Review of Economic Studies* 68: 523–542.
Cooley, T., R. Marimon, and V. Quadrini. 2004. Aggregate consequences of limited contract enforceability. *Journal of Political Economy* 112: 817–847.
Cordoba, J.-C. 2005. *US inequality: Debt constraints or incomplete markets?*, Working paper. Rice University.
Farhi, E., and I. Werning. 2005. *Inequality, social discounting and estate taxation*, Working paper No. 11408. Cambridge, MA: NBER.
Golosov, M., N. Kocherlakota, and A. Tsyvinski. 2003. Optimal indirect and capital taxation. *Review of Economic Studies* 70: 569–587.
Green, E., and S. Oh. 1991. Contracts, constraints and consumption. *Review of Economic Studies* 58: 883–899.
Kehoe, P., and F. Perri. 2002. International business cycles with endogenously incomplete markets. *Econometrica* 70: 907–928.
Kocherlakota, N. 1996. Implications of efficient risk sharing without commitment. *Review of Economic Studies* 63: 595–609.
Kocherlakota, N. 2005. Zero expected wealth taxes: A Mirrleesian approach to dynamic optimal taxation. *Econometrica* 73: 1587–1622.
Kocherlakota, N., and L. Pistaferri. 2005. *Asset pricing implications of Pareto optimality with private information*, Discussion paper No. 4930. London: CEPR.
Lustig, H. 2005. *The market price of aggregate risk and the wealth distribution*, Working paper No. 11132. Cambridge, MA: NBER.
Sleet, C. 2004. Optimal taxation with private government information. *Review of Economic Studies* 71: 1217–1239.
Sleet, C., and S. Yeltekin. 2006a. Optimal taxation with endogenously incomplete markets. *Journal of Economic Theory* 127: 36–73.
Sleet, C., and S. Yeltekin. 2006b. Credibility and endogenous societal discounting. *Review of Economic Dynamics* 9: 410–437.

Energy Economics

Robin Sickles and Hillard G. Huntington

Abstract

Energy economics studies energy resources and energy commodities. It includes forces motivating firms and consumers to supply, convert, transport, use energy resource; market and regulatory structures; distributional and environmental consequences; economically efficient use. The fact that energy use is dominantly depletable resources, particularly fossil fuels, makes this study unique. The energy industry has moved into the 21st century with promises of both profits and a short-term future. With added pressure from government, cleaner fuels are being introduced on a continual basis. Additionally, the expanding energy demand from developing countries is changing the energy market.

Keywords

Conservation; Depletable resources; Derived demand; Dynamic models; Ecological economics; Energy economics; Energy policies; Environmental economics; Essential goods; Framework Convention on Climate Change; Intertemporal choices; Kyoto protocol; Oil; Organization of Petroleum Exporting Countries; Renewable resources; Strategic petroleum reserve (USA)

JEL Classifications

Q4

Energy is crucial to the economic progress and social development of nations.

Energy can be neither created nor destroyed but its form can be changed. Energy comes from the physical environment and ultimately returns there. The demand for energy is a derived demand. The value of energy is assessed by its ability to provide a set of desired services in both industry and in the household.

Energy commodities are economic substitutes. Energy resources are depletable or renewable and storable or non-storable. On a global scale the 20th century was dominated by the use of fossil fuels. According to the US Department of Energy, in the year 2000 global commercial energy consumption consisted of petroleum (39 per cent), coal (24 per cent), natural gas (23 per cent), hydro (6 per cent), nuclear (7 per cent) and others (1 per cent). In 1999, of the total sources of energy

consumed in the United States, 92 per cent were from depletable resources and only 8 per cent from renewable resources (EIA 2001). No one doubts that fossil fuels are subject to depletion, and that depletion leads to scarcity, which in turn leads to higher prices. Resources are defined as 'non-conventional' when they cannot be produced economically at today's prices and with today's technology. With higher prices, however, the gap between conventional and non-conventional oil resources narrows. Ultimately, a combination of escalating prices and technological enhancements can transform the non-conventional into the conventional. Much of the pessimism about oil resources has been focused entirely on conventional resources.

Demand for Energy

Bohi and Toman (1996) suggest a link between energy and economy. An abundance of empirical research suggests a strong correlation between increases in oil prices and decreases in macroeconomic performance for oil-importing industrialized countries. Higher import costs may lead to higher price levels and inflation.

Industrial energy demand increases most rapidly at the initial stages of development, but growth slows steadily throughout the industrialization process (Medlock and Soligo 2001). Energy demand for transportation rises steadily, and takes the major share of total energy use at the latter stages of developments.

Elasticity of Energy Demand

Is energy an essential good? In economics, an essential good is one for which the demand remains positive no matter how high its price. Energy is often described as an essential good because human activity would be impossible absent use of energy. Although energy is essential to humans, neither particular energy commodities nor any purchased energy commodities are essential goods because consumers can convert one form of energy into another.

The income elasticity of energy demand is defined as the percentage change in energy demand given a one per cent change in income holding all else constant, or

$$\varepsilon_y = \frac{\%\Delta e}{\%\Delta y} = \frac{de}{dy} \cdot \frac{y}{e}$$

where e denotes energy demand and y denotes income. 'The household sector's share of aggregate energy consumption tends to fall with income, the share of transportation tends to rise, and the share of industry follows an inverse-U pattern' (Judson et al. 1999).

The price elasticity of energy demand is defined as the percentage change in energy demand given a one per cent change in price, with all else held constant, or

$$\varepsilon_p = \frac{\%\Delta e}{\%\Delta p} = \frac{de}{dp} \cdot \frac{p}{e}$$

where p denotes the price of energy.

Cooper (2003) uses a multiple regression model derived from an adaptation of Nerlove's (1958) partial adjustment model to estimate both the short-run and the long-run elasticity of demand for crude oil in 23 countries over a 30-year period from 1971 to 2000. The estimates so obtained confirm that the demand for crude oil internationally is highly insensitive to changes in price.

Demand Substitution Between Energy Commodities and Others

Denny et al. (1981) used time-series data for 18 US manufacturing two-digit industries (1948–1971) and 18 Canadian manufacturing industry groups (1962–1975). Their results were also mixed: for both the United States and Canada, energy and capital were substitutes in the food industry, but they were complements in the tobacco industry.

Energy consumption can be modelled either as providing utility to households or as an input in the production process for firms. To express the former problem mathematically, a representative consumer maximizes utility, $U(z,e)$, which is function of energy consumption, e, and all other

consumption, z, subject to the constraint that expenditures cannot exceed income, y. Let the energy variable be a vector of n energy products, $e = (e_1, e_2, \ldots, e_n)$; we could examine the substitution possibilities across energy products. Allowing the price of good j to be represented as pj, the consumer is assumed to

$$\max_{z, e_1, \ldots e_n} U(z, e_1, \ldots e_n)$$

$$\text{subject to} : y \geq p_z Z + p_{e_1} e_1 + \ldots + p_{e_n} e_n.$$

The first order necessary conditions for a maximum for this problem can be solved to yield demand equations for each of the energy products and for all other consumption. With some adjustments, the above method can be applied to a representative firm.

Recent research focuses mainly on dynamic models. Dynamic models allow for a more complete analysis of the energy demand because they are capable of capturing factors that generate the asymmetries. In addition, dynamic models incorporate the intertemporal choices that a consumer/firm must make when maximizing utilities or profits over some time horizon. Medlock and Soligo (2002) developed a useful framework. Let z_t be multiple types of capital and e_t be multiple types of energy consumption. Denoting time using the subscript t, the consumer will maximize the discounted sum of lifetime utility, $\sum_{t=0}^{T} \beta^t U(z_t, e_t)$, subject to the constraint whereby capital goods purchases (i_t), purchases of other goods (z_t), purchases of energy (e_t), and savings (s_t) in each period cannot exceed this period's income (y_t), plus the return of last period's saving $((1 + r)s_{t-1})$. It is assumed that capital goods depreciate at a rate δ, savings earn a rate return r, the discount rate is $0 < \beta < 1$, and all initial conditions are given.

Consumers will

$$\max_{z, e, s} \sum_{t=0}^{T} \beta^t U(z_t, e_t)$$

$$\text{subject to } p_{zt} z_t + p_{et} e_t + p_{kt} i_t + s_t \leq y_t + (1 + r) s_{t-1}$$

$$i_t = k_t - (1 - \delta) k_{t-1}$$

$$z_t, u_t, k_t \geq 0 \text{ for } t = 1, \ldots, T$$

Medlock and Soligo (2002) indicate that the income elasticity of passenger vehicle demand is decreasing as the real GDP per capita increases, no matter in the long run or in the short run. For example, with 1988 purchasing power parity dollar, if the real GDP per capita is $500, the short-run elasticity is 0.74 and the long-run elasticity is 3.61; if the real GDP per capita is $20,000, the short-run and the long-run elasticity are 0.02 and 0.09, respectively.

Energy Supply

OPEC

The Organization of the Petroleum Exporting Countries (OPEC) comprises countries that have organized for the purpose of negotiating with oil companies on matters of petroleum production, prices, and future concession rights. Founded on 14 September 1960 at a Baghdad conference, OPEC originally consisted of only five countries – Iran, Iraq, Kuwait, Saudi Arabia and Venezuela – but has since expanded to include several others: Algeria, Indonesia, Libya, Nigeria, Qatar and United Arab Emirates. The members of OPEC, which constitute a cartel, agree on the quantity and the prices of the oil exported. OPEC seeks to regulate oil production, and thereby manage oil prices, primarily by setting quotas for its members. Member countries hold about 75 per cent of the world's oil reserves, and supply 40 per cent of the world's oil. Loury (1990) is an excellent clarification; it studies a dynamic, quantity-setting duopoly game. The author considers a model of competition between two independent firms, A and B, facing indivisibility in production, with given limitations on their cumulative capacities to produce. At date t the flow rates of production of firms A and B are denoted by q_t^a and q_t^b respectively.

The demand side of the market is passively modelled; buyers do not behave strategically. There is an inverse demand function, $P(\cdot)$, which is time invariant and dependent only on the total rate of flow of output of the two firms. Define the discount factor δ^t, a dollar received on date t is worth dollars at date zero. Then their respective payoffs are V_A and V_B where:

$$V_A = \beta \sum_t \delta^t \left[q_t^a p(Q_t)\right]; \text{and} \, V_B = \beta \sum_t \delta^t \left[q_t^b p(Q_t)\right]$$

for β, the lump sum equivalent of the flow of one dollar. It is shown that the ability to precommit can be disadvantageous. Loury (1990) also formalizes the intuition that, when indivisibilities are important, tacit coordination of plans so as to avoid destructive competition is facilitated by establishing a convention of 'taking turns', that is, a self-enforcing norm of mutual, alternate forbearance. Since worldwide oil sales are denominated in US dollars, changes in the value of the dollar against other world currencies affect OPEC's decisions on how much oil to produce. After the introduction of the euro, Iraq unilaterally decided it wanted to be paid for its oil in euros instead of US dollars.

OPEC decisions have a strong influence on international oil prices. A good example is the 1973 energy crisis, in which OPEC refused to ship oil to Western countries that had supported Israel in its conflict with Egypt, the Yom Kippur War. This refusal caused a fourfold increase in oil prices, which lasted five months, starting on 17 October 1973 and ending on 18 March 1974. OPEC nations then agreed, on 7 January 1975, to raise crude oil prices by ten per cent. The high and rising price of oil burdens industrial oil-importing countries in two ways. First, it renders the standard of living lower than otherwise. Second, it affects the economy in ways that are difficult for policymakers to manage: on the one hand, the rising oil price spurs general inflation; on the other hand, it depresses domestic demand and employment. Unlike many other cartels, OPEC has been successful at increasing the price of oil for extended periods. Much of OPEC's success can be attributed to Saudi Arabia's flexibility. It has tolerated cheating on the part of other cartel members, and cut its own production to compensate for other members exceeding their production quotas. This actually gives them good leverage because, with most members at full production, Saudi Arabia is the only member with spare capacity and the ability to increase supply, if needed. The policy has been successful. However, OPEC's ability to raise prices does have some limits. An increase in oil price decreases consumption, and could cause a net decrease in revenue. Furthermore, an extended rise in price could encourage systematic behaviour change, such as alternative energy utilization, or increased conservation. As of August 2004, OPEC has been communicating that its members have little excess pumping capacity, indicating that the cartel is losing influence over crude oil prices.

The six major non-OPEC oil-producing nations are Norway, Russia, Canada, Mexico, the United States and Oman. Russian production increases dominated non- OPEC production growth from 2000 onward and was responsible for most of the non-OPEC increases since the turn of the century. In 2001, a weakening US economy and increases in non-OPEC production put downward pressure on prices.

In response OPEC once again entered into a series of reductions in member quotas, cutting production by 3.5 million barrels per day by 1 September 2001. In the absence of the September 11, 2001 terrorist attack this would have been sufficient to moderate or even reverse the trend.

In the wake of that attack the crude oil price plummeted. Under normal circumstances a drop in price of this magnitude would have resulted in another round of quota reductions, but, given the political climate, OPEC delayed additional cuts until January 2002, when it reduced its quota by 1.5 million barrels per day and was joined by several non-OPEC producers, including Russia, which promised combined daily production cuts of an additional 462,500 barrels. This had the

desired effect, with oil prices moving into the $25 per barrel range by March 2002. By mid-year the non-OPEC members were restoring their production cuts, but prices continued to rise and US inventories reached a 20-year low later in the year. By year's end oversupply was not a problem. Problems in Venezuela led to a strike at Petroleos de Venezuela (PDVSA) causing Venezuelan production to plummet. In the wake of the strike Venezuela was never able to restore capacity to its previous levels. On 19 March 2003, just as some Venezuelan production was beginning to return, military action began in Iraq. Meanwhile, inventories remained low in the United States and other OECD countries. With an improving economy US demand was increasing, and Asian demand for crude oil was growing at a rapid pace. The loss of production capacity in Iraq and Venezuela, combined with increased production to meet growing international demand, led to the erosion of excess oil production capacity. During much of 2004 and 2005 the spare capacity to produce oil has been less than one million barrels per day. A million barrels per day is not enough spare capacity to cover an interruption of supply from almost any OPEC producer. In a world that consumes over 80 million barrels of petroleum products per day, that adds a significant risk premium to crude oil price and is largely responsible for prices in excess of $40 per barrel. For further information, see Energy Information Administration (EIA).

Future Energy Supply

Undoubtedly, depletable resource use cannot dominate forever. Therefore, a future transition from depletable resources, particularly from fossil fuels, is inevitable.

However, which renewable energy sources will dominate future consumption is unclear. And there is great uncertainty about the timing of a shift to renewable energy resources. Although this is a formidable question, Wiser et al. (2004) introduce green pricing programmes, which represent one way whereby consumers can voluntarily support renewable energy. Their analysis yields several interesting results. Programme duration affects customer response. The longer a programme has been operating, the more likely it is that its message has spread and the higher the probability of strong programme success. Initial customer participants in green pricing programmes may not be highly sensitive to cost, and may be willing to purchase higher quantities of renewable energy, which makes the case for utilities focusing on maximizing renewable energy sales, not customer participation rates. Price premiums and minimum monthly costs are not the primary determinants of programme success. Price may become a more important determinant as green pricing programmes expand beyond the early innovator customers. And smaller utilities appear to have a greater likelihood of achieving success.

The prospect of producing clean, sustainable power in substantial quantities from renewable energy sources is arousing renewed interest worldwide. Hydroelectricity is the only renewable energy source today that makes a large contribution to world energy production. Its long-term technical potential is believed to be 9 to 12 times current production, but increasingly environmental concerns block new dams. The large areas affected may have a negative environmental impact. Hydroelectricity dams, like the Aswan Dam, have adverse consequences both upstream and downstream.

Wind power is one of the most cost-competitive renewable sources today. Its long-term technical potential is believed to be five times current global energy consumption. But this requires 12.7 per cent of all land area and the facilities have to be built at certain height. Geothermal power and tidal power are the only renewable sources not dependent on the sun, but are today limited to special locations. Most renewable sources are diffuse and require large land areas and great quantities of construction material for significant energy production. There is some doubt that they can be built rapidly enough to replace fossil fuels. The large and sometimes remote areas may also increase energy loss and cost from distribution. On the other hand, some forms allow small-scale production and may be

placed very close to or directly at consumer households, businesses, and industries. We may forecast the future coexistence of multi-renewable energy sources. Boyle (1996) provides a comprehensive overview of the principal renewable energy sources: solar thermal, biomass, tidal, wave, photovoltaic, hydro, wind and geothermal.

Forecasts of the Energy Markets

According to Energy Information Administration (EIA 2005b), based on its expectations for world energy prices, world energy consumption is projected to increase by 57 per cent from 2002 to 2025. World oil use is expected to grow from 78 million barrels per day in 2002 to 103 million barrels per day in 2015 and 119 million barrels per day in 2025. The projected increment in worldwide oil use would require an increment in world oil production capacity of 42 million barrels per day above 2002 levels.

Members of OPEC are expected to be the major suppliers of the increased production that will be required to meet demand, accounting for 60 per cent of the projected increase in world capacity. In addition, non-OPEC suppliers are expected to add nearly 17 million barrels per day of oil production capacity between 2002 and 2025. Substantial increments in new non-OPEC supply are expected to come from the Caspian Basin, Western Africa, and Central and South America.

Natural gas is projected to be the fastest-growing component of world primary energy consumption. Consumption of natural gas worldwide increases in the forecast by an average of 2.3 per cent annually from 2002 to 2025, compared with projected annual growth rates of 1.9 per cent for oil consumption and 2.0 per cent for coal consumption. From 2002 to 2025, consumption of natural gas is projected to increase by 69 per cent, and its share of total energy consumption is projected to grow from 23 to 25 per cent.

Natural gas is seen as a desirable alternative to electricity generation in many parts of the world, given its relatively efficiency in comparison with other energy sources, as well as the fact that it burns more cleanly than either coal or oil and thus is an attractive alternative for countries pursuing reductions in greenhouse gas emission.

World coal consumption is projected to increase at an average rate of 2.5 per cent per year. From 2015 to 2025, the projected rate of increase in world coal consumption slows to 1.3 per cent annually. Coal is expected to maintain its importance as an energy source in both the electric power and industrial sectors.

Hydroelectricity and other renewable energy sources are expected to maintain their 8 per cent share of total energy use worldwide throughout the projection period. Much of the projected growth in renewable electricity generation is expected to result from the completion of large hydroelectric facilities in emerging economies, particularly in Asia.

Energy Policies

The study of depletable resource economics began with articles by H. Hotelling (1931), which examined economically intertemporal optimal extraction from a perfectly known stock of the resource, with perfectly predictable future prices of the extracted commodity. Sweeney (1977) and Stiglitz (1976) both clarified the Hotelling rule in the presence of monopoly, and Gilbert and Richard (1978) and Salant (1976) extended this to the case of a dominant producer with a competitive fringe and several dominant producers, analogous to the case of OPEC. Pindyck (1982) and Kolstad (1994) extended the model to several imperfectly substitutable exhaustible resources.

Energy security refers to loss of economic welfare that may occur as a result of a change in the price of availability of energy. In the years following the 1973 oil price rise, US energy policy could be characterized as generally suspicious of the market. Supply augmentation was a major strategy pursued by the US government in addressing the ‘energy crisis’. The security dimensions of energy supply have always been viewed as appropriate concerns of the government. One could argue that the Gulf War in the early 1990s was

simply a form of energy policy, protecting Western oil supplies originating in the Middle East. Countries other than the United States (such as Japan and China) have tried to diversify their sources of energy to reduce the risk of disruption. Security was also viewed as threatened by sudden fluctuations in the price of oil, hence the establishment in the United States of the Strategic Petroleum Reserve (SPR): petroleum.

Stocks are maintained by the federal government for use during periods of major supply interruption. The idea is that, if the price of oil were to rise rapidly due to disruption in supply, then the SPR could be called upon to provide supplies, thus reducing the price shock.

Nuclear power was declared dead in the United States because it is too expensive and unacceptably risky. Around the world, nuclear plant ended up achieving less than ten per cent of the new capacity and one per cent of the new orders (all from countries with centrally planned energy systems) forecast in the early 1980. The industry has suffered the greatest collapse of any enterprise in industrial history. Scientists still have not developed reliable ways to handle nuclear wastes and decommissioned plants, which remain dangerously radioactive for far longer than societies last or geological foresight extends.

Strong economic growths across the globe and new global demands for more energy have meant the end of sustained surplus capacity in hydrocarbon fuels and the beginning of capacity limitations. In fact, the world is currently precariously close to utilizing all of its available oil-production capacity, raising the chances of an oil-supply crisis with more substantial consequences than seen since the early 1970. These limits mean that the United States can no longer assume that oil-producing states will provide more oil. Nor is it strategically and politically desirable for the United States to remedy its present tenuous situation by simply increasing its dependence on a few foreign sources. As a result, expanding demand for energy will change US policy towards the Middle East, Russia and China. A recent example is that, in 2005, the state-owned Chinese company CNOOC eventually abandoned its bid for Unocal due to strong political opposition in the United States.

Effects of Energy Demand

Energy and Macroeconomics

In fact, almost every recession since the Second World War in the United States, as well as many other energy-importing nations, has been preceded by a spike in the price of energy (Hamilton 1983; Ferderer 1996; Mork et al. 1994).

The oil price movement affects certain sectors: oil-dependent manufacturing such as paper and packaging, consumer-related sectors such as autos, refiners' margins, the energy-intensive utility sector, and of course exploration companies and the big oil majors themselves.

Energy, Economy and Environment

Many important environment damages stem from the production, conversion, and consumption of energy. The costs of these environmental damages generally are not incorporated into prices for energy commodities and resources; this omission leads to overuse of energy. It has been shown that estimates of damage costs resulting from combustion of fossil fuels, if internalized into the price of the resulting output of electricity, could clearly lead to a number of renewable technologies being financially competitive with generation from coal plants. Environmental impacts currently receiving most attention are associated with the release of greenhouse gases in the atmosphere, primarily carbon dioxide, from the combustion of fossil fuels. During combustion, carbon combines with oxygen to produce carbon dioxide, the primary greenhouse gas. Carbon dioxide accumulates in the atmosphere and is expected to result in significant detrimental impacts on the world's climate, including global warming, rises in the ocean levels, increased intensity of tropical storms, and losses in biodiversity. Concern about this issue is common to energy economics, environmental economics, and ecological economics. Cropper and Oates (1992) suggest measuring benefits and costs with a review of cases where benefit–cost analyses have actually been used in the setting of environmental standards. Owen (2004) suggests that penalizing high pollutant-emitting technologies not only creates incentives

for 'new' technologies but also encourages the adoption of energy-efficiency measures with existing technologies and consequently lower pollutants per unit of output.

World carbon dioxide emissions are expected to increase by 1.9 per cent annually between 2001 and 2025. Much of this increase is expected to occur in developing countries. The United States produces about 25 per cent of global carbon dioxide emissions from burning fossil fuels, primarily because of it has the largest economy in the world and meets 85 per cent of its energy needs through burning fossil fuels. The United States is projected to lower its carbon intensity by 25 per cent from 2001 to 2025. There are numerous proposals aimed at reducing the carbon dioxide emissions, of which the Kyoto Protocol is a well-known and influential one. During 1–11 December 1997, more than 160 nations met in Kyoto, Japan, to negotiate binding limitations on greenhouse gases for the developed nations, pursuant to the objectives of the Framework Convention on Climate Change of 1992. The outcome of the meeting was the Kyoto Protocol, in which the developed nations agreed to limit their greenhouse gas emissions relative to the levels emitted in 1990. The United States agreed to reduce emissions from 1990 levels by seven per cent during the period 2008 to 2012.

Sickles and Jeon (2004) evaluate the role that undesirable outputs of the economy, such as carbon dioxide and other greenhouse gases, play on the frontier production process. This paper also explores implications for growth of total factor productivity in the OECD and Asian economies.

Natural disasters shock the energy market, too. According to the Minerals Management Service (2005), Gulf of Mexico daily oil production was reduced by 89 per cent as a result of Hurricane Katrina in 2005. The MMS also reports that 72 per cent of daily Gulf of Mexico natural gas production was shut in. In 2004, Hurricane Ivan caused lasting damage to the energy infrastructure in the Gulf of Mexico and interrupted oil supplies to the United States. US Secretary of Energy Spencer Abraham agreed to release 1.7 million barrels of oil in the form of a loan from the Strategic Petroleum Reserve.

A Concluding Comment

The world runs on energy, primarily energy generated from coal and petroleum. The current war against terrorism and the tensions in the Middle East have raised new questions about the reliability of America's oil supply from that region. Concerns about global climate change have also focused increased attention on the search for cleaner fuels and energy-generating methods. Russia's determination to become a major petroleum supplier, OPEC's periodic moves to restrict oil production and the rising energy needs in China and other developing countries are all important issues forming the future world energy market.

I would like to thank Robert Thomure, Rice University, for his research assistance.

See Also

- ▶ Environmental Economics
- ▶ Oil and the Macroeconomy
- ▶ Organization of the Petroleum Exporting Countries (OPEC)

Bibliography

Bohi, D., and M. Toman. 1996. *The economics of energy security*. Boston: Kluwer Academic Publishers.

Boyle, G. 1996. *Renewable energy: Power for a sustainable future*. Oxford: Oxford University Press.

Cooper, J. 2003. Price elasticity of demand for crude oil: Estimates for 23 countries. *OPEC Review: Energy Economics & Related Issues* 27 (1): 1–8.

Cropper, M., and W. Oates. 1992. Environmental economics: A survey. *Journal of Economic Literature* 30: 675–740.

Denny, M., M. Fuss, and L. Waverman. 1981. Substitution possibilities for energy: Evidence from U.S. and Canadian manufacturing. In *Modeling and measuring natural resource substitution*, ed. E. Berndt and B. Field. Cambridge, MA: MIT Press.

EIA (Energy Information Administration). 2001. *Annual energy review 2000*. Washington, DC: EIA.

EIA. 2004. *Annual energy review 2004*. Washington, DC: EIA.

EIA. 2005a. *Annual energy review 2005*. Washington, DC: EIA.

EIA. 2005b. *Annual energy outlook 2005 with projections to 2025*. Washington, DC: EIA.

Fang, F., and R. Sickles. 2004. The role of environmental factors in growth accounting. *Journal of Applied Econometrics* 19: 567–591.
Ferderer, P. 1996. Oil price volatility and macroeconomy. *Journal of Macroeconomics* 18: 1–26.
Gilbert, R., and J. Richard. 1978. Dominant firm pricing policy in a market for an exhaustible resource. *Bell Journal of Economics* 9: 385–395.
Hamilton, J. 1983. Oil and the macroeconomy since World War II. *Journal of Political Economy* 91: 228–248.
Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.
Judson, R., R. Schmalensee, and T. Stoker. 1999. Economic development and the structure of the demand for commercial energy. *Energy Journal* 20 (2): 29–27.
Kolstad, C. 1994. Hotelling rents in hotelling space: Product differentiation in exhaustible resource markets. *Journal of Environmental Economics and Management* 26: 163–180.
Loury, G. 1990. Tacit collusion in a dynamic duopoly with indivisible production and cumulative capacity constraints. Working Paper No. 557. Department of Economics, MIT.
Medlock, K., and R. Soligo. 2001. Economic development and end-use energy demand. *Energy Journal* 22 (2): 77–105.
Medlock, K., and R. Soligo. 2002. Car ownership and economic development with forecasts to 2015. *Journal of Transport Economics and Policy* 36: 163–188.
MMS (Minerals Management Service). 2005. *Hurricane information*. Washington, DC: MMS.
Mork, K., H. Mysen, and O. Olsen. 1994. Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy Journal* 15 (4): 19–35.
Nerlove, M. 1958. Distributed lags and demand analysis for agricultural and other commodities. In *Agriculture handbook*, No. 141. Washington, DC: US Department of Agriculture.
Owen, A. 2004. Environmental externalities, market distortions and the economics of renewable energy technologies. *Energy Journal* 25 (3): 127–156.
Pindyck, R. 1982. Jointly produced exhaustible resources. *Journal of Environmental Economics and Management* 9: 291–303.
Salant, S. 1976. Exhaustible resources and industrial structure – Nash–Cournot approach to the world oil market. *Journal of Political Economy* 84: 1079–1093.
Sickles, R.B., and B.M. Jeon. 2004. The role of environmental factors in growth accounting. *Journal of Applied Econometrics* 19: 567–591.
Stiglitz, J. 1976. Monopoly and rate of extraction of exhaustible resources. *American Economic Review* 66: 655–661.
Sweeney, J. 1977. Economics of depletable resources – Market forces and intertemporal bias. *Review of Economic Studies* 44: 125–141.
Wiser, R., S. Olson, L. Bird, and B. Swezey. 2004. *Utility green pricing programs: A statistical analysis of program effectiveness.* Berkeley: Ernest Orlando Lawrence Berkeley National Laboratory and National Renewable Energy Laboratory. Online. Available at http://www.eere.energy.gov/greenpower/resources/pdfs/lbnl_54437.pdf. Accessed 1 Aug 2006.

Energy Price Shocks

Lutz Kilian

Abstract

Oil price shocks have been a recurring phenomenon since the 1970s. This article reviews alternative explanations of oil price shocks. It puts the evolution of the US price of crude oil into historical perspective and compares it with that of the price of coal and natural gas.

Keywords

Oil; Coal; Natural gas; History; Shocks; Price determinants

JEL Classifications

Q43

Energy is necessary to sustain a country's real economic activity and to ensure the physical survival of its population, except under the most favorable climatic conditions. The main sources of energy to this day have been fossil fuels, such as coal, crude oil and natural gas. Electricity in turn is produced mainly from power plants burning coal or natural gas (or, in rare cases, fuel oil), augmented by nuclear power and to a lesser extent by hydroelectric, solar and wind power.

The prices of oil, coal and natural gas thus are of immediate concern to policymakers, firms and consumers. Unexpected changes in the price of energy have the potential to wreak havoc on an economy. Unanticipated increases in energy prices may cause far-reaching disruptions of economic plans, as consumers and firms are forced to economise on the use of energy, to curtail other expenditures to pay for higher energy costs and to

replace energy-inefficient equipment with energy-saving equipment. Such energy price shocks and their effects on the economy are the subject of a large literature in economics.

Much of this literature has focused on the price of crude oil (see, for example, Barsky and Kilian (2002), Kilian (2008a, 2014) and Hamilton (1983, 2003, 2008)). Crude oil stands out among energy commodities because of its importance for the transportation sector. Although crude oil is not consumed directly and is not used as a factor of production outside of the refining industry, oil products such as petrol/ gasoline, diesel, heating oil or jet fuel are highly visible in everyday life. Their prices affect, for example, the cost of commuting to work, the cost of shipping goods, the cost of air travel and in some areas the cost of home heating.

Until the early 1970s, the global market for crude oil looked much different from typical industrial commodity markets, with the USA able to produce most of the oil it consumed and regulating the price of domestically produced crude oil, while other industrialised countries, such as Japan or Germany, were heavily dependent on crude oil imports. This situation changed when even the USA became heavily dependent on oil imports in the early 1970s, following an increase in domestic oil demand. Initially the required additional oil imports came primarily from the Middle East, but later increasingly from oil producers in other regions as well. As global oil trade expanded and national oil market structures were gradually broken up, a global market for crude oil emerged. From 1974 until 2010, the price of crude oil has been determined in this global market place. As a result, adjusting for transportation costs and the inevitable differences in the quality of crude oil produced in different regions of the world, the price of crude has been largely the same worldwide. In recent years, as the production of unconventional crude oil surged in the USA and Canada, the global market for crude oil appears to have fragmented, with US oil prices deviating from world prices because of bottlenecks in transporting crude oil and because of capacity constraints in refining (see Kilian 2015).

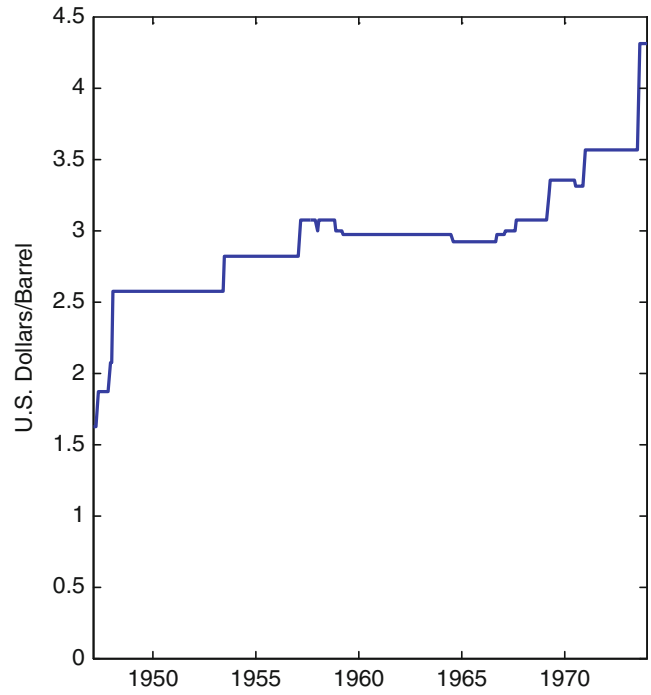
Long before the emergence of a global market for crude oil in the 1970s, there was already a well-developed global market for coal. Well into the 20th century, coal was as important for the US transportation sector as oil is today. It was the primary fuel used in shipping until the 1920s and in railroading until the 1950s. It also served as an important source of home heating until the 1950s. Coal continues to play an important role today in producing electricity and heat as well as in the manufacturing of chemicals and metals.

In producing electricity and in manufacturing coal competes with natural gas. The market for natural gas, in contrast to that of coal or crude oil, has largely remained regional to this day. Natural gas is primarily transported by pipelines. Although natural gas may be cooled down and liquefied, allowing it to be shipped as liquefied natural gas (LNG) to any port in the world, both the cost of LNG shipping and the infrastructure required to load and unload LNG are expensive, which has prevented the integration of regional natural gas markets and the emergence of a global price to date.

Thus, in studying energy prices over longer time periods one inevitably has to take the perspective of one country. This article focuses on the USA because of its long history in producing coal, oil and natural gas and the availability of long continuous price series for these markets. The objective is to document the history of energy price shocks and to examine whether price shocks in oil markets are different from those in other energy markets such as coal, which in the past were as important as crude oil is today.

The Traditional Interpretation of Oil Price Shocks

It is useful to begin with a review of the oil price data in the post-Second World War period. Sometimes oil price shocks are associated with sudden increases in the price of oil. This idea can be traced back to the work of Hamilton (1983) who studied the evolution of the nominal price of West Texas Intermediate (WTI) crude oil in the USA between 1948 and 1972. Figure 1 illustrates that

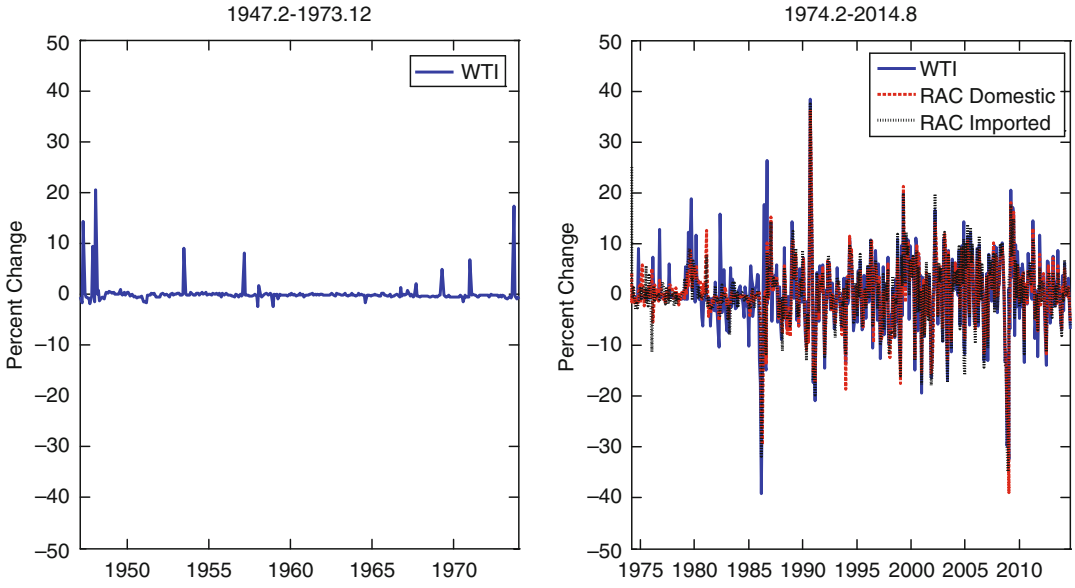


Energy Price Shocks, Fig. 1 Evolution of the monthly WTI price of crude oil during 1947.1–1973.12 (the WTI data are from the US Energy Information Administration (EIA))

this price series differs from most other commodity prices in that it often remains unchanged for extended periods, followed by discrete adjustments. When expressed in growth rates and adjusted for inflation, as shown in the left panel of Fig. 2, this step-function pattern implies unpredictable spikes in the growth rate that represent shocks to the price of oil.

This pattern reflects the fact that the US price of crude oil during this period was regulated by state-level agencies such as the Texas Railroad Commission, as discussed in Hamilton (1983). Under normal circumstances the regulator was able to keep the price of oil unchanged for extended periods. At irregular intervals, however, major oil price adjustments took place. Hamilton documents that these adjustments were associated with oil supply disruptions in the Middle East that were unforeseen by the regulator and that justified ex-post adjustments of the regulated price of oil. These events could be used either to implement long overdue adjustments to the price of oil reflecting domestic inflation and/or unexpectedly high domestic demand for oil, or to accommodate additional demand for US oil from abroad. In short, Hamilton's review of the evidence suggested that the timing of the oil price shocks under the Texas Railroad Commission regime was associated with oil supply shocks in the Middle East driven by political events unrelated to the state of the US economy, allowing us to treat them as exogenous with respect to the US economy.

This regime came to an end in the early 1970s, when demand for oil grew much faster than US oil production and the US economy became heavily dependent on oil imports. Much of the additional oil was imported from the Middle East. The fact that state-level agencies were unable to regulate the price of imported crude oil spelled the end of the Texas Railroad Commission regime, even though the last vestiges of this system survived until the early 1980s. After 1974, the market for crude oil, for all intents and purposes, became a global market with the price of oil being ultimately determined by the forces of demand and supply, much as in other global commodity markets. This structural shift in the crude oil market is reflected in a structural break in the evolution of the growth rate of the real price. Figure 2 shows a



Energy Price Shocks, Fig. 2 Evolution of the growth rate of the monthly real price of oil (the data are from the US EIA and Federal Reserve Economic Data (FRED). RAC denotes the US refiners' acquisition cost)

dramatic increase in the volatility of three alternative measures of the growth rate of the price of oil, but especially of the US refiners' acquisition cost for crude oil imports, which may be viewed as a proxy for the global price of crude oil. Rather than exhibiting occasional spikes, the growth rate of the real price of oil as of 1974 begins to look like that of most other commodity prices.

The Modern Interpretation of Oil Price Shocks

Initially, it was thought that this structural change was inconsequential and that at least the major fluctuations in the real price of crude oil after 1973 could be explained by exogenous oil supply disruptions abroad, much like those in the 1950s and 1960s. Kilian (2008b) demonstrated that this is not the case. Examples of political events in oil-producing countries thought to have triggered oil supply disruptions include the 1973 Yom Kippur War and the subsequent Arab oil embargo, the Iranian Revolution of 1978–79, the invasion of Kuwait in 1990, the Venezuelan unrest of late 2002 and the Iraq War of early 2003, as well as the Libyan Revolution of 2011. The challenge for the traditional view of oil price shocks has been that the predictive power of these supply disruptions for the price of oil is quite modest. Oil supply disruptions explain at most a quarter of the increase of the price of oil in 1973–74, for example, and with the exception of the 1990 spike in the level of oil price have not had a major impact on the evolution of the real price of oil since 1974. The major oil price fluctuations instead appear to be driven by shifts in the demand for oil, as has been shown in a series of studies, including Baumeister and Peersman (2013), Kilian (2009), Kilian and Hicks (2013), and Kilian and Murphy (2012, 2014).

By far the most important determinant of the real price of oil is shifts in the flow demand for oil associated with fluctuations in the global business cycle. Flow demand refers to demand for raw materials to be consumed right away in the process of producing more domestic goods rather than being stored for future use. As China's industrial growth accelerates unexpectedly, for example, the flow demand for industrial raw materials, including crude oil, increases. As the demand curve shifts to the right along the upward-sloping

supply curve, the real price of crude oil (and of other industrial raw materials) increases. Put differently, the real price of oil is endogenous with respect to global macroeconomic conditions. This phenomenon attracted much attention after 2003, but is by no means new. Shifts in the flow demand for oil played a major role during almost all major surges in the real price of oil including the 1973–74 and 1978–80 episodes.

Another potentially important determinant of the real price of oil is shifts in the demand for oil stocks, reflecting forward-looking behaviour by oil market participants. Such demand shocks are also known as speculative demand shocks. They arise, for example, when market participants expect the real price of oil to go up in the future, reflecting expectations of a shortfall of future supply relative to future demand. In this case there is an incentive to buy crude oil now and to store it in anticipation of rising oil prices. The resulting shift in the current demand for oil stocks increases the current real price of oil, as the demand curve shifts to the right along the supply curve.

Such forward-looking behaviour is crucial for understanding oil markets. It has been shown that exogenous political events in the Middle East matter for the real price of oil not so much because of the actual disruptions of the flow of crude oil they cause, but because of the expectations of future supply disruptions that they may create. Likewise, the anticipation of a global economic recovery or economic slowdown will affect expectations of future oil prices, as will any number of other events and developments that are not commonly included in economic models. Even an increase in uncertainty, all else equal, may cause a shift in the demand for oil stocks (see Pindyck 2004; Alquist and Kilian 2010). As Kilian and Murphy (2014) and Kilian and Lee (2014) show, speculative demand shocks driven by expectations of future oil price changes help explain, for example, the surge in the real price of oil in 1979 (following the Iranian Revolution), the collapse of the real price of oil in 1986 (following the collapse of OPEC) and the spike in the real price of oil in 1990 (following the invasion of Kuwait), but they played no important role during the 2003–08 surge in the real price of oil.

Finally, there are a myriad additional idiosyncratic shocks to the demand for oil, ranging from politically motivated changes to the Strategic Petroleum Reserve before elections to shifts in the demand for oil as a result of hurricanes in the Gulf of Mexico shutting down US refineries. These idiosyncratic demand shocks, however, do not appear to be capable of explaining sustained changes in the real price of oil.

Once it is recognised that not all oil price shocks are the same, it becomes immediately clear that one would expect the evolution of the US economy in the wake of an oil price shock to differ depending on the composition of the oil demand and oil supply shocks underlying this price shock. If we ignore this insight, we may find that the statistical relationship between the real price of oil and the US economy appears unstable over time, even when the underlying structural relationship is stable. This point has been illustrated, for example, by Kilian and Park (2009) in the context of US stock markets.

One can still ask how the US economy responds on average to an oil price shock, of course, but there are two important caveats in interpreting the answer. First, these responses cannot be interpreted as the causal effects of the oil price shock, because nothing ensures that the price shock under consideration occurs holding everything else constant. For example, an unexpected increase in the real price of oil driven by increased flow demand would also be associated with increases in the real prices of other industrial raw materials, violating the *ceteris paribus* assumption. Thus the response we observe in the economy is the response to increases in the prices of both oil and industrial raw materials, rather than the price of oil alone.

Second, the average response to an oil price shock can be misleading when it comes to interpreting specific episodes of rising oil prices. For example, traditional models of oil price shocks implied that the US economy should have gone into recession in 2005–06, following the surge in the price of oil that began in 2003. Such a recession obviously never occurred, because the preceding increase in oil prices was caused by an unexpectedly booming global

economy, not by oil supply disruptions or increased speculative demand. Unlike in the traditional view of oil price shocks as being driven entirely by exogenous oil supply disruptions, in the modern view, rising oil prices may very well be compatible with an expanding economy and a rising stock market, at least for some time.

A final point to bear in mind is that the traditional question of how an oil price shock affects the economy becomes inherently ill-posed once we recognise that the state of the economy in turn affects the price of oil. A more useful way of posing this question would be to ask how an exogenous shift in the demand for oil in some part of the world, for example, affects the real price of all commodities including crude oil, the US economy and the economy in rest of the world. Answering the latter question requires a global structural model of the economy and of commodity markets. An example of this type of analysis can be found in Bodenstein et al. (2012). Related work also includes Nakov and Pescatori (2010).

Other Explanations of Oil Price Shocks

Especially near the peak of the real price of oil in 2008, a popular view among some pundits has been that the real price of oil is no longer determined by the laws of demand and supply, but by the actions of financial traders in oil futures markets (sometimes informally referred to as financial speculators). This view reflects several misunderstandings. One is that it is logically impossible for the price of crude oil to be determined by anything else but demand or supply. The only question is what determines the demand for and supply of crude oil. In fact, economic theory suggests that prices in the physical market for crude oil and in the oil futures market are jointly and simultaneously determined by the same underlying shocks rather than changes in one price being caused by exogenous changes in the other. Another misunderstanding is that the actions of financial speculators in the oil futures market are believed to be exogenous with respect to developments in the physical market for oil. Not only is it unclear what exactly a financial speculator is, but the claim of exogenous financial speculation moving oil markets is difficult to sustain in practice. Proponents of this view have not been able to provide convincing empirical evidence in support of the financial speculation hypothesis. For a review of this debate the reader is referred to Fattouh et al. (2013).

Upon closer examination, it becomes clear that the real concern articulated by these pundits is not about a failure of the laws of demand and supply at all, but about the perception that economic fundamentals, as measured by shocks to the flow of oil being produced and shocks to the flow demand for oil, are not capable of explaining the surge in the real price of oil, especially in 2007–08. This is a misperception. Formal empirical models show that economic fundamentals do an excellent job at explaining the surge in the price of oil between 2003 and mid-2008, as well as the collapse and recovery of the real price of oil thereafter (see, for example, Kilian and Murphy (2014) and Kilian and Lee (2014)).

The debate about financial speculation as a driver of oil prices illustrates a tendency among pundits to reduce complicated economic relationships in oil markets to simple formulaic explanations. The notion of nefarious speculators in oil markets is one example of trying to make sense of the evolution of the real price of oil without engaging with economic models. Another example is the tendency to attribute oil price increases to actions of the so-called OPEC cartel. OPEC is the Organization of Petroleum Exporting Countries. It includes oil producers in the Middle East as well as some oil producers in other parts of the world. The presumption is that there is a concerted effort by OPEC oil producers to prop up oil prices either directly or by withholding oil supplies from the market. Until ten years ago, major oil price increases were routinely attributed to the machinations of this alleged cartel rather than to the underlying forces of demand and supply. Only with the rise of the debate about financial speculation does interest among pundits in OPEC seem to have waned.

The evidence in support of the cartel explanation has always been thin (e.g. Smith 2005;

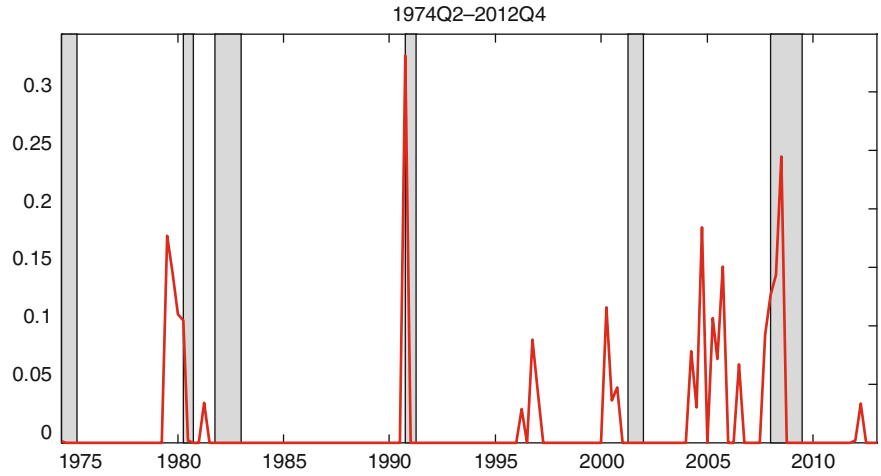
Almoguera et al. 2011). There is no evidence that OPEC caused the 1973–74 or the 1979–80 oil price shock episodes, for example, even if this claim is often repeated in macroeconomic textbooks. OPEC was far from a unified body in the 1970s and incapable of acting as a cartel. Only in the early 1980s did OPEC attempt to curtail its oil production in an effort to prevent oil prices from falling in response to the Volcker recession. As predicted by the economic theory of cartels, most OPEC members cheated on their cartel obligations, however, which prompted Saudi Arabia to take responsibility for reducing oil production on behalf of the rest of OPEC. This approach proved not only ineffective, in that the price of oil continued to fall (albeit at a slower rate), but unsustainable in that falling production in conjunction with falling oil prices resulted in a substantial reduction in Saudi oil revenues. By late 1985, Saudi Arabia was forced to reverse course, and the real price of oil collapsed along with fears of what OPEC might do. There has been no indication of OPEC being able or willing to control the price of crude oil since then. Indeed, it would be hard to explain why Saudi Arabia would have permitted the oil price to fall to $11/barrel in 1998 if it were endowed with the market power sometimes ascribed to it.

How to Measure Oil Price Shocks

So far we have defined an oil price shock informally as a change in the price of oil relative to the price of oil that consumers and firms expected. More formally, oil price shocks can be defined as the unpredictable component of the price of oil. One approach is to measure oil price shocks within the context of an econometric model as movements in the price of oil that cannot be explained based on past data. Such oil price shocks are also known as oil price innovations, and can be decomposed further into mutually uncorrelated oil demand and oil supply shocks with the help of additional identifying assumptions. An alternative approach would be to define oil price shocks based on the market expectation of the price of oil. For example, Baumeister and Kilian (2014) discuss how oil price expectations for a given horizon may be recovered by adjusting the oil futures price by an empirical measure of the risk premium. By comparing the three-month ahead financial market expectations of the oil price to the realisations of the oil price three months later, for example, one can infer a time series of quarterly oil price shocks.

More colloquially, the term ‘oil price shock’ is also used to denote episodes of unusually high (or in some cases unusually low) oil prices. Such episodes typically extend over several years. In fact, most surges in the price of oil do not involve any large changes in the price of oil on a monthly basis. Rather, they arise because for extended periods the price of oil experiences small but persistent increments. Examples are the 1979–80 and 2003–08 oil price surges. Sometimes appearances can be misleading. A case in point is the sudden increase in the price of oil in 1973–74. Kilian (2008b) shows that the real price of oil would have increased much earlier than late 1973, and more gradually, had the price of Middle Eastern oil not been constrained by the Tehran-Tripoli agreement of 1971. The sudden increase in late 1973 occurred when oil producers reneged on these contractual agreements and the oil price reverted to market levels. In fact, overall the real price of metals and non-oil industrial raw materials between 1971.11 and 1974.2 increased by 75% as much as the real price of oil, even in the absence of supply shocks in these markets, suggesting that all these prices were largely driven by the same forces of demand. Hence, historically, the oil price spike of 1990, following the invasion of Kuwait, is the only example of a large and sudden increase in the price of oil since the 1960s. All other oil price shock episodes have involved more gradual increases in the real price of oil.

Finally, yet another notion of oil price shocks has been proposed by Hamilton (1996, 2003). The idea is that an oil price shock occurs only to the extent that the price of oil exceeds the highest price of oil that consumers and firms have experienced in recent memory. More formally, this *net oil price increase* measure of oil price shocks is defined as the censored variable max (0, $pt-p^*$),



Energy Price Shocks, Fig. 3 Three-year net oil price increase in real US refiners' acquisition cost for oil imports with US recession dates imposed as shaded areas (Source: Kilian and Vigfusson (2014). The business cycle dates are from the National Bureau of Economic Research)

where *pt* refers to the current oil price and *p** refers to the maximum oil price over the preceding year (or, more commonly, over the preceding three years). By construction, the net oil price increase is predictable based on its own past. At some point it was believed that this statistical transformation of the price of oil would effectively isolate the component of the price of oil associated with exogenous shocks to the flow supply of oil. It has become readily apparent that this is not the case. An alternative and more common interpretation has been that net oil price increases are the relevant measure of oil price shocks because they explain or at least help predict variation in US real GDP.

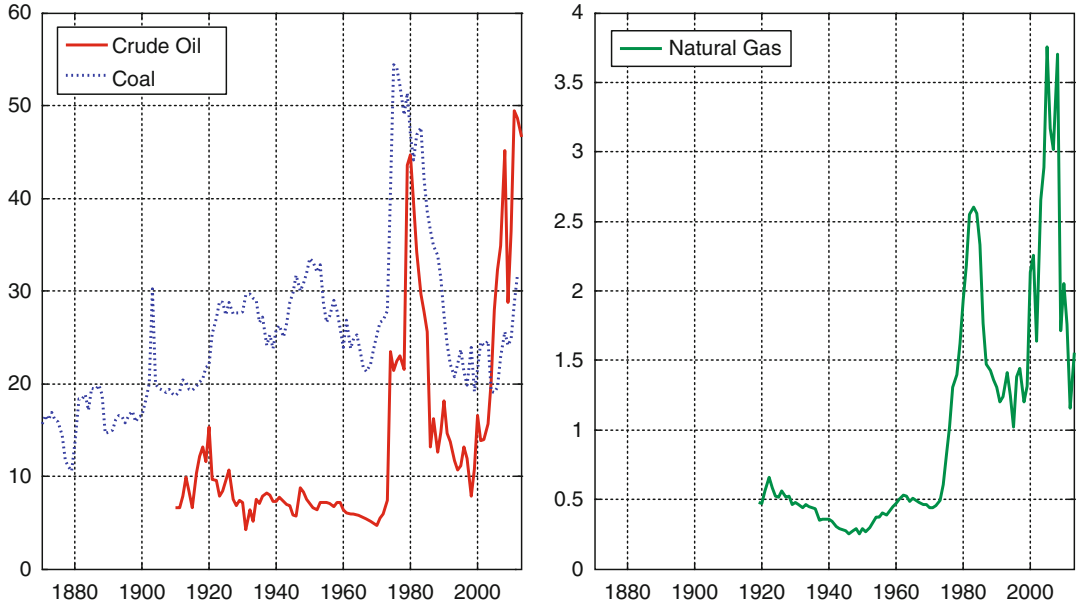
Figure 3 casts doubt on this interpretation. Treating the net oil price increases of 2004–06 as one episode, there have been eight distinct episodes of net oil price increases since 1974, of which only five were followed by recessions. In some cases the net oil price increase occurred well before the recession. A good example is the net oil price increase of 2000. In other cases it occurred immediately before or at the same time as the recession. Examples are 1981 and 1990. In three cases, net oil price increases were not followed by a recession at all. These episodes are 1996 as well as 2004–05 and 2006 (the latter two may be viewed as one episode), and 2011–12. This evidence suggests that there is no mechanical link between net oil price increases and subsequent recessions.

More formally, it can be shown that the evidence that net oil price increases help forecast US real GDP growth is weak at best (Ravazzolo and Rothman 2013; Kilian and Vigfusson 2013). For related discussion of net oil price increase measures and their relationship with more conventional oil price shock measures, see Kilian and Vigfusson (2011a, b).

Putting Oil Price Shocks into Historical Perspective

Crude oil is only one source of primary energy, but it stands out because of its important role in the transportation sector. Historically, coal played much the same role for the transportation sector as oil did starting with the increased adoption of the automobile during the First World War. Steam ships and steam locomotives were as dominant in transportation then as oil is today for trucking, shipping, air transport and railroading. A natural question therefore is whether coal prices were subject to shocks similar to the oil price shocks documented earlier.

This question can only be addressed with annual data, because there are no quarterly or

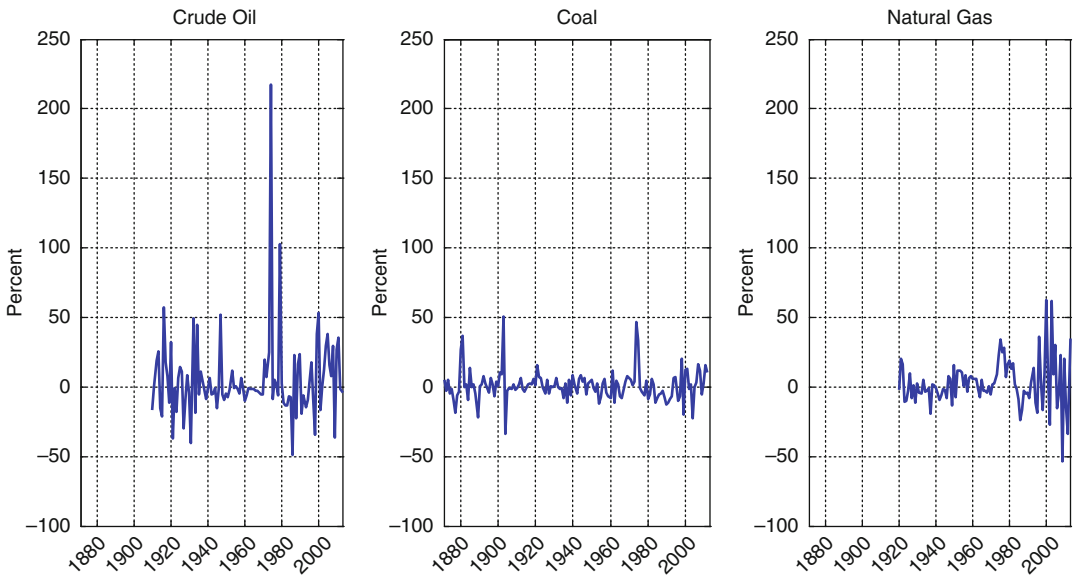


Energy Price Shocks, Fig. 4 Historical evolution of US real energy prices: 1870–2013 (expressed in real US$ per barrel (crude oil), per metric ton (coal) and per MMBtu (natural gas). The coal price refers to anthracite coal. The coal and gas prices are based on Manthy (1978) and EIA sources; the oil price series is from British Petroleum (2014))

monthly oil price series prior to 1947. The left panel of Fig. 4 plots the historical evolution of the real prices of coal and crude oil since 1870. The plot deliberately ignores the oil price data prior to 1910, because, before the adoption of the automobile, crude oil was used primarily to produce kerosene to be used for lighting, heating and cooking. It therefore makes sense to discount the history of oil prices prior to 1910. Figure 4 shows that prior to 1970 the degree of comovement between the prices of oil and coal was quite low. Subsequently, there is increased comovement, but the increase in the real price of coal during the 2000s was not nearly as dramatic as that in the real price of oil.

Because these prices are measured in different units, it is useful to express them in percentage changes. Figure 5 suggests that the Texas Railroad Commission era, which was characterised by unusually low volatility at annual frequency followed by an extreme spike in 1973–74, was a historical aberration. In contrast, the volatility of the real price of oil prior to the Second World War largely resembles that since the 1970s. Even discounting the early 1970s, however, the volatility in the growth rate of the annual price of oil appears to be more than twice as high as the corresponding volatility in the price of coal. This is not to say that there are no sustained increases in the real price of coal – in fact the sustained increases in the level of the real price of coal during the 1920s and 1940s dwarfed those in crude oil – but that the year-on-year changes tended to be smaller. In this sense, there is a clear difference between the crude oil market and the coal market.

It is also instructive to compare the real price of oil with the real wellhead price of natural gas. Although natural gas prices are available as far back as 1919, as shown in the right panel of Fig. 4, it was only with the creation of a nationwide network of natural gas pipelines in the 1950s that natural gas became an important energy resource for the US economy (see Davis and Kilian 2011). Figure 5 shows that the volatility of the growth rate of the real price of natural gas since the late 1950s has been quite similar to that of the real price of oil before and after the Texas



Energy Price Shocks, Fig. 5 Percent growth rates of the real US price of energy: 1871–2013 (all results are obtained from the data underlying Fig. 4)

Railroad Commission interlude. Figure 4 in turn suggests that traditionally US natural gas and crude oil prices have moved in the same direction. There are indications, however, that with the rapid growth in US shale gas production in recent years this traditional pattern no longer holds. Figure 4 shows a dramatic fall in the real price of natural gas after 2008, even as the annual real price of oil recovered after the financial crisis. In contrast, there remains some degree of positive comovement between coal and crude oil prices going back as far as the 1970s.

The observed pattern of positive comovement across coal, oil and natural gas prices between the 1970s and the 2000s reflects in part the fact that industrial consumers of energy often had the ability to use dual technologies that allowed them to switch between natural gas and fuel oil, depending on price and availability. Such substitution tends to be more difficult for residential consumers of energy, however. For example, to this day there are parts of the USA in which heating oil remains the main source of home heating because there are no natural gas pipelines in that region. Likewise, there is essentially no substitution between oil and either coal or natural gas in US power plants, and the process of replacing coal power plants by natural gas power plants is quite slow. Finally, especially in transportation, there is only very limited substitutability between oil and natural gas to date. Even the indirect substitution of coal, natural gas or nuclear power for oil in the form of electric power has not played a large role in US transportation so far.

Thus, much of the observed comovement in real energy prices appears to reflect common shifts in the demand for all forms of primary energy associated with shifts in flow demand. This comovement only breaks down when there are large increases in the supply of an energy commodity, as occurred in the natural gas sector after 2008. In contrast, there is no indication that a similar structural shift is under way in the oil market. At the global level, the quantitative importance of US shale oil remains small compared with the size of the market for crude oil. Thus, the response of the global price of oil to the US shale oil revolution has been muted (see Kilian 2015). In contrast, in the natural gas sector, US gas production must be balanced against domestic demand rather than global demand, with correspondingly larger effects on the price.

An interesting question for future research will be to compare the evolution of energy prices across countries and to disentangle the contributions of various demand and supply shifts to the evolution of these historical energy price series. The latter question has received increasing scrutiny in years, including contributions by Hamilton (2013), van de Ven and Fouquet (2014), and Stürmer (2014). For very long run trends in energy prices such as those of charcoal, coal, town gas and kerosene see Fouquet (2011).

Conclusions

Why do we care about oil price shocks (and by extension other energy price shocks)? One reason is that positive oil price shocks historically have been associated with recessions in oil-importing countries, although the recessionary effects associated with oil price shocks do not appear to be as large and systematic as originally thought.

Of course, not all increases in the real price of oil are bad. Increases in the price of oil may also serve to transmit signals about the increased scarcity of crude oil (see Hamilton 2014). In fact, rising oil prices are a precondition for the development and adoption of alternative energy technologies. In this sense, the concern for policymakers is not so much increases in the real price of oil in general, but rather the fact that rapid increases in the real price of oil tend to put economic stress on the oil-importing economy as the economy adjusts to the increased scarcity of oil.

An even bigger concern is high volatility in the growth rate of the real price of oil, which may prevent the necessary investment in alternative energy technologies or for that matter in additional oil exploration (see Dixit and Pindyck 1994; Kellogg 2014). There has been no shortage of discussions of the need to stabilise oil prices. One response has been the creation of the US Strategic Petroleum Reserve (SPR). It is clear, however, that relative to the magnitude of the global oil market, changes in the SPR are too small to stabilise the real price of oil. Of course, the biggest source of volatility in oil prices in recent years has been the financial crisis. The case can be made that policies preventing such misalignments in the economy may be the most effective approach to reducing the volatility of oil prices and other primary energy prices.

See Also

- ▶ Business Politics in the Gulf
- ▶ Energy Economics
- ▶ Oil and Politics in the Gulf: Kuwait and Qatar
- ▶ Oil and the Macroeconomy
- ▶ Oman, Economy of
- ▶ Organization of the Petroleum Exporting Countries (OPEC)
- ▶ Yemen, Economy of

Acknowledgments I thank Martin Stürmer for providing access to the annual energy price data used in this article. I have also benefitted from helpful discussions with Christiane Baumeister, Ryan Kellogg and Roger Fouquet, and the comments of a referee.

Bibliography

Almoguera, P.A., C.C. Douglas, and A.M. Herrera. 2011. Testing for the cartel in OPEC: Noncooperative collusion or just noncooperative? *Oxford Review of Economic Policy* 27: 144–168.

Alquist, R., and L. Kilian. 2010. What do we learn from the price of crude oil futures? *Journal of Applied Econometrics* 25: 539–573.

Barsky, R.B., and L. Kilian. 2002. Do we really know that oil caused the great stagflation? A monetary alternative. In *NBER macroeconomics annual 2001*, ed. B.S. Bernanke and K.S. Rogoff, Vol. 16, 137–183. Cambridge: MIT Press.

Baumeister, C., and L. Kilian. 2014. *A general approach to recovering market expectations from futures prices with an application to crude oil*. Mimeo: University of Michigan.

Baumeister, C., and G. Peersman. 2013. The role of time-varying price elasticities in accounting for volatility changes in the crude oil market. *Journal of Applied Econometrics* 28: 1087–1109.

Bodenstein, M., L. Guerrieri, and L. Kilian. 2012. Monetary policy responses to oil price fluctuations. *IMF Economic Review* 60: 470–504.

British Petroleum. 2014. Statistical review of world energy, June.

Davis, L.W., and L. Kilian. 2011. The allocative cost of price ceilings in the U.S. residential market for natural gas. *Journal of Political Economy* 119: 212–241.

Dixit, A.K., and R.S. Pindyck. 1994. *Investment under uncertainty*. Princeton: Princeton University Press.
Fattouh, B., L. Kilian, and L. Mahadeva. 2013. The role of speculation in oil markets: What have we learned so far? *Energy Journal* 34: 7–33.
Fouquet, R. 2011. Divergences in long run trends in the prices of energy and energy services. *Review of Environmental Economics and Policy* 5: 196–218.
Hamilton, J.D. 1983. Oil and the macroeconomy since World War II. *Journal of Political Economy* 91: 228–248.
Hamilton, J.D. 1996. This is what happened to the oil price–macroeconomy relationship. *Journal of Monetary Economics* 38: 215–220.
Hamilton, J.D. 2003. What is an oil shock? *Journal of Econometrics* 113: 363–398.
Hamilton, J.D. 2008. Oil and the macroeconomy. In *The new Palgrave dictionary of economics*, 2nd ed., ed. S. Durlauf and L. Blume. Basingstoke: Palgrave MacMillan.
Hamilton, J.D. 2013. Oil prices, exhaustible resources and economic growth. In *Handbook of energy and climate change*, ed. R. Fouquet, 29–57. Cheltenham: Edward Elgar.
Hamilton, J.D. 2014. The changing face of world oil markets. Mimeo, UC San Diego.
Kellogg, R. 2014. The effect of uncertainty on investment: Evidence from Texas oil drilling. *American Economic Review* 104: 1698–1734.
Kilian, L. 2008a. The economic effects of energy price shocks. *Journal of Economic Literature* 46: 871–909.
Kilian, L. 2008b. Exogenous oil supply shocks: How big are they and how much do they matter for the U.S. economy? *Review of Economics and Statistics* 90: 216–240.
Kilian, L. 2009. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review* 99: 1053–1069.
Kilian, L. 2014. Oil price shocks: Causes and consequences. *Annual Review of Resource Economics* 6: 133–154.
Kilian, L. 2015. *The impact of the shale oil revolution on U.S. oil and gas prices*. Mimeo, University of Michigan.
Kilian, L., and B. Hicks. 2013. Did unexpectedly strong economic growth cause the oil price shock of 2003–2008? *Journal of Forecasting* 32: 385–394.
Kilian, L., and T.K. Lee. 2014. Quantifying the speculative component in the real price of oil: The role of global oil inventories. *Journal of International Money and Finance* 42: 71–87.
Kilian, L., and D. Murphy. 2012. Why agnostic sign restrictions are not enough: Understanding the dynamics of oil market VAR models. *Journal of the European Economic Association* 10: 1166–1188.
Kilian, L., and D. Murphy. 2014. The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics* 29: 454–478.
Kilian, L., and C. Park. 2009. The impact of oil price shocks on the U.S. stock market. *International Economic Review* 50: 1267–1287.
Kilian, L., and R.J. Vigfusson. 2011a. Are the responses of the U.S. economy asymmetric in energy price increases and decreases? *Quantitative Economics* 2: 419–453.
Kilian, L., and R.J. Vigfusson. 2011b. Nonlinearities in the oil price–output relationship. *Macroeconomic Dynamics* 15: 337–363.
Kilian, L., and R.J. Vigfusson. 2013. Do oil prices help forecast U.S. real GDP? The role of nonlinearities and asymmetries. *Journal of Business and Economic Statistics* 31: 78–93.
Kilian, L. and R.J. Vigfusson 2014. The role of oil price shocks in causing U.S. Recessions. Mimeo, University of Michigan.
Manthy, R.S. 1978. *Natural resource commodities – A century of statistics*. Baltimore: Johns Hopkins Press for Resources for the Future.
Nakov, A., and A. Pescatori. 2010. Monetary policy tradeoffs with a dominant oil producer. *Journal of Money, Credit, and Banking* 42: 1–32.
Pindyck, R.S. 2004. Volatility and commodity price dynamics. *Journal of Futures Markets* 24: 1029–1047.
Ravazzolo, F., and P. Rothman. 2013. Oil and U.S. GDP: A real time out-of-sample examination. *Journal of Money, Credit & Banking* 45: 449–463.
Smith, J. 2005. Inscrutable OPEC? Behavioral tests of the cartel hypothesis. *Energy Journal* 26: 51–82.
Stürmer, M. 2014. 150 years of boom and bust: What drives mineral commodity prices? Mimeo, Federal Reserve Bank of Dallas.
van de Ven, F.J., and R. Fouquet 2014. *Historical energy price shocks and their changing effects on the economy*. Centre for climate change economics and policy working paper No. 171.

Energy Services

Roger Fouquet

Abstract

Energy consumers are driven by their demand for energy services (such as space and water heating, cooking, transportation, lighting, entertainment and computing). This piece introduces the reader to the concept of energy services, and explains why it is important to analyze energy markets and climate policies from the perspective of energy services. The

paper discusses the theoretical foundations and empirical evidence, particularly related to the rebound effect and the demand in developing economies. The paper concludes that governments should encourage the collection of statistical information about energy services in order to help economists analyse markets and policies through this lens. Most importantly, governments should formulate more integrated policies that focus explicitly on energy services, connecting markets for energy and for energy-using equipment with the development of technologies. Careful and balanced energy service policies are especially important as economies industrialise because they can help reduce economic, political and environmental vulnerability.

Keywords

Energy consumption; Energy services; Energy policy; Climate change; Consumer behaviour; Direct rebound; Price elasticity

JEL Classifications

Q32; Q38; Q43; Q48; Q58

What Are 'Energy Services'?

Energy services refer to the services that are generated from consuming energy combined with appliances. For residential consumers, these services include space heating and cooling, water heating, cooking, refrigeration, lighting, computing, entertainment and passenger transport (Reister and Devine 1979, 1981; Goldemberg et al. 1985). For firms, these include high temperature processes, such as iron smelting, low temperature processes, moving of motors and machinery, separation, drying, and compressing air, as well as refrigeration, lighting, space and water heating, and freight transport.

For space heating and cooling, the service is measured in terms of the increase or decrease in temperature compared with the existing temperature (in degrees Centigrade or Farenheit) for a specified surface area. For example, a 100 square metre house is warmed 10°C for two hours – which is the equivalent to warming 2,000 square metres 1°C for one hour. Transport is measured in terms of passenger-kilometres (or –miles), or tonne-kilometres (or ton-miles) for freight. The more efficient the vehicle, then the more passenger-kilometres can be achieved with, say, a litre or gallon of gasoline. The number of vehicle-kilometres in a country, divided by the gasoline consumed in that country, offers an indicator of the average fuel efficiency of vehicles (Frondel et al. 2008; Stapleton et al. 2016). For lighting, the unit of measurement is lumen-hours, which indicates the amount of illumination generated by a light source (Nordhaus 1997). A 100 watt incandescent bulb provides about 1,300 lumens. So, if it is left on for 10 hours, it will have generated 13,000 lumen-hours and used 1,000 watts. The same amount of lighting could have been produced over the same amount of time using a 20 watt CFL bulb, but consuming only 200 watts.

Energy services are closely related to end-use energy consumption and the concept of exergy. End-use energy consumption refers to the energy consumed for individual services. However, focusing on end-use energy consumption or prices does not take account of the efficiency of conversion of the energy into the service. Exergy, on the other hand, does take account of the efficiency and refers to the amount of 'work' produced (Ayres and Warr 2009). The strength of the concept of exergy is that it looks at all energy services in the same unit; however, it does not focus on the nature of the specific output, which risks ignoring important dimensions of the analysis (Sovacool 2011).

The purpose of this piece is to introduce readers to the concept of energy services. The next section explains why energy economists are increasingly analyzing energy service consumption, rather than only energy use. The third section presents the foundations for analyzing the demand and provision of energy services. In the fourth section, empirical evidence of the demand for energy services in the residential and transport sectors is presented, focusing on price elasticities and the size of direct rebound effects. The fifth

section discusses the two-way relationship between energy services and economic development. Due to space constraints, other topics related to energy services are not discussed in detail – including fuel poverty (Boardman 2010), exergy (Ayres and Warr 2009), the impact of smart meters (Römer et al. 2012) and of net-metering (Gillingham et al. 2016a), energy service companies or ESCOs (Weiller and Pollitt 2013), and energy systems planning and operation, such as demand-side management (Strbac 2008). The final section draws conclusions about our knowledge of energy services and its role in improving policy-making.

The Importance of Energy Services

Energy consumption is driven by the demand for energy services. Individuals do not consume electricity for the voltage that speeds down the wires, or put gasoline in their cars for the pleasure of having a full tank. Instead, they consume them because of the lighting the electricity creates or the mobility the gasoline provides.

Consequently, studies of energy demand may strongly benefit from considering the relationship between energy services, technologies and energy consumption (Haas et al. 2008). If the relationship remains constant, it may not be crucial to focus explicitly on energy services. For instance, in the short run, the relationship does stay broadly constant, because the efficiency of the technology (that is, the amount of service generated for a given unit of energy) does not change much and, so, the focus on energy services may not alter the results of the analysis.

However, analyzing energy services is especially important in the long run, as technological change can radically alter energy consumption behaviour. Nordhaus (1997) showed how the price of fuel for lighting fell three-fold and the price of lighting (measured in lumen-hours) fell 75-fold in the last century, thus, traditional methods of measuring the price of lighting using the fuel price were off by a factor of 25. Fouquet (2011) and Muller (2016) showed that this is not an isolated example and that, in general, the trend in the (nominal and real) price of an energy service diverges from the trend in the price of energy for this service in the long run. While the average energy price has not shown a discernible trend in the long run, the price of energy services has tended to fall. The divergence implies that focusing exclusively on energy prices and consumption rather than energy services will generate misleading conclusions about energy consumption behavior. This long run perspective is particularly relevant when thinking about climate change mitigation.

As Fouquet and Pearson (2012) argued, by not focusing on energy services, the analyst is making an implicit assumption about the price elasticity of demand for the energy service. Two 'straw-man' examples can be used to show this. First, the 'efficiency optimist' might suggest that if energy efficiency improves by 10%, energy consumption will fall by 10%. However, since the efficiency has improved by 10%, the consumer can get the same quantity of service with 10% less energy. This implies that the price of the energy service has fallen 10%. For energy consumption to fall by 10%, energy service use must remain unchanged. So, the 'efficiency optimist' implicitly assumes that the price elasticity of demand for energy services is zero. Alternatively, the 'laggard economist' might propose that since the price of energy is unchanged consumption of energy will remain the same. In this case, since the price of this energy service has fallen by 10%, for energy consumption to remain unchanged, energy service use must increase by 10%. So, the implicit assumption here is that the price elasticity of demand for energy services is one. Thus, focusing on energy rather than energy services forces the analyst to make assumptions about consumer behavior and is likely to create misleading estimates of consumer responses to long run energy price and efficiency changes. Ultimately, the size of the price elasticity of demand is an empirical question and needs to be estimated in order to help identify the scale of the 'rebound effect', which will be discussed in the fourth section.

Hunt and Ryan (2015) have made this point explicit, emphasizing the misspecification of models that fail to incorporate energy service

demand and the biased elasticity estimates that result. In their analysis, the income elasticity of demand for energy is underestimated and the price elasticity is overestimated, because of the failure to model energy services and include energy efficiency improvements. However, they explain that the bias depends on the trends in income, real prices and efficiency improvements, implying that it is not possible to generalize the direction of bias (Hunt and Ryan 2015, p. 283).

An additional advantage of focusing on energy services is that the demand for energy services stays relatively stable with the introduction of new energy sources and technologies. Traditional analysis sees energy transitions as disruptive events – with a radically declining demand for, say, biomass fuels and rapidly rising demand for coal – with no continuity. However, they can be seen as competing technologies and sources for producing the same energy service. In this way, long run patterns in energy service consumption can be identified that would be hidden by focusing only on the uptake and decline of energy sources and technologies (Fouquet 2014).

In addition, Smulders and de Nooij (2003) highlight the limitations of a study of economic growth that ignores energy services. They show that, within their model, energy conservation policies, which reduce energy consumption, lower economic growth. However, this assumes that growth in energy use is a key source of the growth in economic output, rather than energy service consumption, which is unlikely to decline following energy conservation policies. Indeed, Toman and Jemelkova (2003) emphasize the importance of energy services in driving economic development. They show that energy services can affect economic development through a number of different channels, and that these effects can change at different levels of economic development, and it is essential to model them explicitly.

Finally, the Nordhaus (1997) piece sought to highlight that the consumer price index (CPI) and the gross domestic product (GDP) are mismeasured if they do not take account of increases in the quality of service provision, which result from technological improvements. Lighting is just one example amongst many in which conversion of a good into a service is underestimated. In fact, Nordhaus (1997, p. 60) suggests that 'estimates of the growth of real consumption services is hampered by significant errors in the measurement of prices and that for almost two-fifths of consumption the price indexes are virtually useless.'

In other words, focusing on energy services rather than energy consumption can greatly improve our understanding of energy consumption behavior, including the rebound effect (see the fourth section), of the relationship between energy markets and economic growth, and even of fundamental measurements of cost-of-living and economic activity. The main reason economists have tended to ignore energy services has been a lack of data on energy efficiency to convert data into services. As Sorrell (2007, p. 25) explains: 'For many energy services, the relevant data is simply unavailable, while for others the data must be either estimated or subject to considerable error.'

The Demand for Energy Services and Its Household Production

Having discussed the importance of focusing on energy services, and before reviewing the empirical evidence, it is valuable to outline briefly the basic theory underlying the demand and provision of energy services. Energy service markets often involve the same agent demanding and providing the service by consuming energy and acquiring related equipment (that is, the physical capital). Here, the focus is on the residential and transport sector, although similar issues apply to energy service markets in industrial and tertiary sectors. One difference is the increasing separation of demand and supply with ESCOs (Energy Service Companies) providing the services, which will be briefly discussed.

The first modelling of the derived demand for energy, combining complementary durable equipment, dates back to Houthakker (1951). Early studies highlighted the fundamental importance of the relationship between energy use and appliances, but were not explicit about the consumer's objectives related to energy services (Berndt and

Wood 1975; Pindyck 1979; Hausman 1979; Khazzoom 1980; Dubin and McFadden 1984; Dubin et al. 1986) – for a review of the early literature on energy demand modelling, see Taylor (1975). Then, a growing literature emphasized the importance of modelling energy end-use or service consumption, starting with Reister and Devine (1979, 1981), Neels (1981), Goldemberg et al. (1985), Quigley (1984), Klein (1988), and Quigley and Rubinfeld (1989), though focusing on the production of services.

However, economists have been slow to explicitly model the demand for energy services, and were eventually stimulated by Nordhaus' (1997) seminal piece on the price of lighting, by Modi et al.'s (2006) emphasis on the provision of energy services in developing economies and by the interest in the rebound effects that hamper efforts to mitigate climate change through energy efficiency improvements. The following outline summarises the demand-side perspective presented in Hunt and Ryan (2015), while incorporating the supply-side approach proposed by Neels (1981) and Quigley (1984), which is also discussed in Frondel et al. (2008).

A consumer or household's objective is to maximize utility – here, the focus is explicitly on taking account of the energy services consumption (ES) generated for meeting this utility:

$$\text{Max } U_t = u(ES_t, X_t), \tag{1}$$

subject to constraints

$$Y_t = P_{ESt} \cdot ES_t + P_{Xt} \cdot X_t \tag{2}$$

where X_t is a composite of goods and services, P_{ESt} and P_{xt} refer to the prices of the energy services and of the composite goods, and Y_t is the consumer's budget, which should be permanent wealth (although it is often proxied by income). Other constraints, for example, relating to the availability of information, technical problems using certain products and the existence of institutional factors which influence the ability to make decisions and to choose goods, might also be included for a more realistic (but more complicated) optimization problem.

Based on the above analysis, but for simplicity assuming only economic constraints, utility depends indirectly on prices and income; the indirect utility function is

$$U'_t = (P_{ESt}, P_{xt}, Y_t). \tag{3}$$

The fact that the indirect utility function represents the consumption of energy services and composite goods as a function of prices and income is particularly valuable for analyzing economic behaviour since neither utility nor preferences can be observed, whereas prices and income can. Thus, for example, the demand function for energy services is

$$ES_t = f(P_{ESt}, P_{xt}, Y_t). \tag{4}$$

By specifying the nature of the optimization problem, principally the constraints faced by consumers, and solving it, we can examine the way optimal choices vary with changing constraints. Tracing out these variations in consumption, the behavioural relationships between consumption and constraints, such as described in the energy service demand function, can be identified. Knowledge of the energy service demand function, for example, can then be used to assess the implications of changing economic activity and policies on fuel consumption. The effects of these changing constraints can be analyzed in the form of the own price elasticity of demand:

$$\varepsilon_{PESt} = (\partial ES_t/ES_t)/(\partial P_{ESt}/P_{ESt}), \tag{5}$$

the income elasticity of demand:

$$\eta_{Yt} = (\partial ES_t/ES_t)/(\partial Y_t/Y_t), \tag{6}$$

and the cross price elasticities:

$$\varepsilon_{Pxt} = (\partial ES_t/ES_t)/(\partial P_{xt}/P_{xt}). \tag{7}$$

This conventional model of consumer behavior outlines the demand for energy services. The supply of energy services is less conventional, however. Rooted in Becker's (1965) theory of the allocation of time, households produce their

own services by combining labour, capital and energy. At a larger scale, firms similarly generally produce their own energy services.

Technological developments over the last two centuries have led to a move away from labour inputs and towards physical capital and energy sources, implying that many energy services, such as heating and lighting, are now provided with virtually no labour requirements. Car driving is the only energy service where substantial labour is required today – and suggests that the diffusion of driver-less cars, and the associated decline in the labour costs (in time), may have a significant impact on the consumption of passenger transport services. With this feature of the modern provision of most energy services, a simplified model of the household production would include only capital (k_t) and energy used (e_t).

The relationship depends on the efficiency of the technology (ϕ_t) – that is, the amount of energy services generated by a specified quantity of energy. As Hunt and Ryan (2015, p. 274) explain: 'three particular characteristics of energy-using equipment are of relevance: much of it is longlived – once installed it may have a useful life that spans decades; much of it is fuel(s)-specific; and its technical characteristics tend to be fixed, requiring a given level of energy use per unit of services produced.' Given that this relationship is often a constant at any point in time, the provision of energy services can be determined by the energy consumption multiplied by the efficiency of the appliance:

$$ES_t = \varphi_{et} \cdot e_t. \tag{8}$$

Frondel et al. (2008) highlight that improvements in energy efficiency may be associated with higher capital costs. Therefore, ideally, the cost of producing energy services should take account of an estimate of these capital costs, as well as any time expenditure and the price of energy. However, a common assumption made is that the price of energy services is determined by the marginal cost of production, which is generally simplified to the price of energy (P_{et}) divided by the technical efficiency of the appliance being used (see Nordhaus 1997):

$$P_{ESt} = P_{et}/\varphi_{et}. \tag{9}$$

Feeding Eqs. 8 and 9 into Eqs. 5 and 6 enable energy economists to estimate the own-price and income elasticity of demand for energy services.

This section presented a simple model of the demand for energy services in which the consumer also produced the service. This implies that the same consumer and producer is actively involved in selecting the production technology and the energy sources. Recently, energy service companies (ESCOs) have begun to take on the responsibility for producing these services. While this can create a principal-agent problem, it is also seen as a way to stimulate energy efficiency improvements and reduce the energy efficiency gap (Gillingham and Palmer 2014). If these companies expand their role beyond the provision to firms, in the future, energy service markets may become more conventional, in the sense of the consumer and producer being different agents - for more on this particular development, see Groscurth et al. (1995), Olerup (1998), and Weiller and Pollitt (2013).

The opposite may be occurring in the market for power. While the final services associated with power (for example some heating, cooling, lighting, entertainment, computing, and so on) are provided by consumers, they have not produced their own power since the early days of electricity generation. However, since the introduction of micro-wind turbines and the drop in the price of solar panels, more households are becoming 'prosumers'. That is, consumers are entering the market for the production of electricity (and, in some cases, selling their surplus, known as 'net-metering'), and blurring the roles (Römer et al. 2012; Gillingham et al. 2016a). In other words, no single model can capture the different characteristics of all energy service consumption and provision. Nevertheless, the model presented above outlines a simple framework for thinking about the market for energy services.

The Direct Rebound Effect and the Price Elasticity of Demand for Energy Services

The main reason energy services have received a great deal of attention in the last decade is due to the debate about rebound effects. They refer to consumer, producer and market responses to energy efficiency improvements (Sorrell and Dimitropoulos 2008; Gillingham 2014). As mentioned before in the second section, they include a direct effect on the consumption of energy services and, thus, energy in response to a higher efficiency improvement and lower energy service price. There are also indirect effects on consumption behaviour related to complements and substitutes of the cheaper energy service (and associated energy source), to a probable increase in purchasing power (after taking account of the expenditure on the new efficient technology) and, therefore, to an increase in the consumption of other goods and services. Finally, macroeconomic rebound effects occur because the reduction in the price of energy services tends to boost the economy, stimulating further energy service and energy consumption. Thus, for instance, a 10% improvement in energy efficiency is unlikely to lead to a 10% saving in energy use. Instead, the sizes of the different and combined rebound effects are empirical questions (see, for instance, Sorrell 2007; Gillingham 2014).

Despite the recent interest, the origins of the debate on the size of the rebound effects began 150 years ago. In 1865, William Stanley Jevons published *The Coal Question*. As a leading political economist of the time, his book sought to shed light on the murky debates surrounding the potential exhaustion of coal resources that were central to Britain's economic supremacy (Madureira 2012). One of his most controversial passages in the book warned that '....it is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption. The very contrary is the truth.... Every improvement of the engine when effected will only accelerate anew the consumption of coal...' (1865). The idea that energy efficiency improvements could lead to increases in energy consumption became known as Jevons' Paradox.

Jevons' Paradox (also now known as 'backfire') is effectively an extreme case in which the rebound effects are sufficiently large that the efficiency improvements lead to increases in consumption. There is now a large theoretical literature supporting the existence of rebound effects which either implicitly or explicitly analyze the price elasticity of demand for energy services (Khazzoom 1980; Saunders 1992; Howarth 1997; Turner 2013; Gillingham and Chan 2015). However, empirical studies have tended to estimate much smaller rebound effects than Jevons (1865) anticipated. So, in the recent cases investigated, energy efficiency improvements led to savings in energy consumption, all other things being equal (Greening et al. 2000; Sorrell 2009). Thus, the inconsistency between Jevons' predictions and the recent empirical evidence suggests a paradox to the Jevons' Paradox.

Ultimately, measuring all the different (i.e. the direct, indirect and macroeconomic) rebound effects empirically at the same time is challenging (Gillingham 2014; Gillingham et al. 2016b). "Measuring the rebound effect is not an easy task, as it involves an estimation of the elasticity of the demand for a particular energy service with respect to energy efficiency. Instead of using this original definition, the majority of available studies have estimated the rebound effect using price elasticity, since data on energy efficiency has always been limited. In principle, rational consumers should respond in the same way to a decrease in energy prices as they do to an improvement in energy efficiency. This assumption, however, does not always hold up, as energy efficiency itself may be affected by changes in energy prices" (Sorrell 2007, p. 4).

Nevertheless, the price elasticity of demand for energy services offers a means of estimating the direct rebound effect associated with efficiency improvements. As Hunt and Ryan (2015) explain, a number of early studies tried to include data on energy efficiency, either by using a deterministic or a stochastic trend (Beenstock and Willcocks

1981; Dimitropoulos et al. 2005) or by measuring energy efficiency directly or indirectly (Walker and Wirl 1993; Haas and Schipper 1998; Haas and Biermayr 2000; Fouquet and Pearson 2012; Fouquet 2014; Schleich et al. 2014). Earlier studies used the efficiency indicator as an additional explanatory variable. The more recent studies used these measures of efficiency to produce indicators of the price and consumption of energy services, which were used to estimate the price elasticity of demand for energy services.

Frondel et al. (2008) outline the assumptions made in efforts to estimate this price elasticity and direct rebound effects. Ideally, as explained in the third section, the price elasticity of demand for energy services can be estimated based on variations in the fixed costs of capital (and labour, associated with the capital investment), and the marginal costs of labour and energy services. However, this is rarely done or even possible, and a second-best is to estimate the elasticity based on variations in the marginal cost of energy services – as a number of the later studies above did. These studies ignore the endogeneity of the fixed costs of capital and the marginal cost of the energy services (as often more efficient equipment is more expensive). Finally, traditional studies have used the price elasticity of demand for energy as a proxy for energy services. Frondel et al. (2008) offer a rare study where all three methods were used on the same data, and so provide an opportunity to compare the results. The authors were surprised to find that the price elasticity estimates using the three different methods were similar, but the coefficients on other explanatory variables were substantially different. Thus, their study highlights the ambiguity of using only energy data given that consumer behaviour is driven by energy service demand.

Given the greater availability of data on transport use and energy consumption related to transport services, this service has been studied most extensively and has offered an opportunity to estimate actual price elasticities of energy services and measure the direct rebound effect. For instance, using a panel data set of US states between 1960 and 2004, Small and van der Dender (2007) estimated the long run price elasticity of demand for car transport to be −0.22 in the second half of the twentieth century, falling to −0.06 between 2001 and 2004. This implies that the direct rebound effect associated with a 10% efficiency improvement fell from 2.2 to 0.6%. Focusing on the more expensive and densely populated Great Britain, Stapleton et al. (2016) estimated the direct rebound effects for car transport over a similar time period to have ranged from 0.9 to 3.6%. The similar results for these two studies suggest that the widely different economic, political and behavioural characteristics may not have influenced greatly the sensitivity to changes in the price of car transport. On the other hand, Frondel et al. (2008) found substantially larger direct rebound effects for Germany – averaging 5.8% for a 10% efficiency improvement, which they explain as due to greater potential for substitution between modes of transport.

While some uncertainty about the scale of the direct rebound effect still remains, the growing number of studies are offering a range of values for the price elasticities of demand for various energy services. The first effort to summarise the finding was in Greening et al. (2000), indicating the range to be between 0 to −0.5, with a concentration in the range of −0.1 to −0.3. More recent efforts include Sorrell (2007), Azevedo (2014), Gillingham (2014), and Gillingham et al. (2016b). The latter selected estimates from nine studies based on rigorous identification strategies, and argued that this lowers slightly the range (between −0.05 and −0.40). An early example of a randomized controlled trial (that is, an experiment set up purposefully to identify the causality) associated with energy efficiency improvements found that the price elasticity of demand for clothes washing was −0.06 (Davis 2008). Table 1 presents estimates for a few key energy services based on a general review of the literature. The broad conclusion is that direct rebound effects are an important issue, but they are unlikely to lead to Jevons' Paradox (or 'backfire') for households or personal transport in developed countries – without drawing a conclusion about the combined impact of direct, indirect and macroeconomic rebound effects – see

Chitnis and Sorrell (2015) for an attempt to measure the combined effects.

As discussed earlier, modelling energy service demand is important for explaining past behaviour, forecasting future consumption and anticipating the impact of policies, including efforts to mitigate climate change and potentially begin the transition towards a low carbon energy sources (Pearson 2016). An important issue is the projection of dramatic increases in air conditioning demand and use over the next few decades, because of declining costs of air conditioning and electricity, improving energy efficiency, and rising incomes and temperatures in developing economies, with potential positive feedback loops (Davis and Gertler 2015). Other studies, such as Anandarajah et al. (2009), Anandarajah and Strachan (2010) and Fujimori et al. (2014), also explicitly model energy service demands for their long run scenarios – see Table 2, as an example. These studies show the relevance of the estimates for practical purposes. However, these are generally based on limited reviews of the evidence, and the assumptions made in the model tend to remain constant through time. Indeed, a key issue raised in the literature reviews, such as Azevedo (2014), Gillingham (2014) and Gillingham et al. (2016b), is about the 'external validity' of the studies. That is, it is unclear whether those estimates will be the same if different methods or models are used and in different time periods or contexts. Gillingham et al. (2016b) emphasize the empirical strategy used, and that these studies tend to assume other characteristics related to the energy source and technology remain unchanged and increases in energy efficiency are costless. Azevedo (2014) stresses that most studies are for the residential and transport sectors in developed economies, particularly in the US.

In fact, over decades, price elasticities of demand for energy services appear to have changed considerably as per capita income has increased (Fouquet 2014). Estimates for residential heating, transport and lighting in the United Kingdom indicate that price elasticities peaked (at values of about −1.5) at levels of per capita income of between $(2010) 4,000 and $(2010) 5,000 (see Fig. 1, bottom-half). That is, in Britain in the 1870s and 1880s, a 10% reduction in energy prices or a 10% improvement in energy efficiency (both reducing the price of energy services) increased transport and lighting use by around 15%. This implies that energy efficiency improvements associated with transport and lighting led to rises in energy consumption, as Jevons (1865) had predicted – offering an explanation for the paradox of Jevons' paradox. Furthermore, given that elasticities of demand for energy services change, efforts should be made to incorporate more

Energy Services, Table 1 Estimates of price elasticities of demand for energy services in industrialised economies

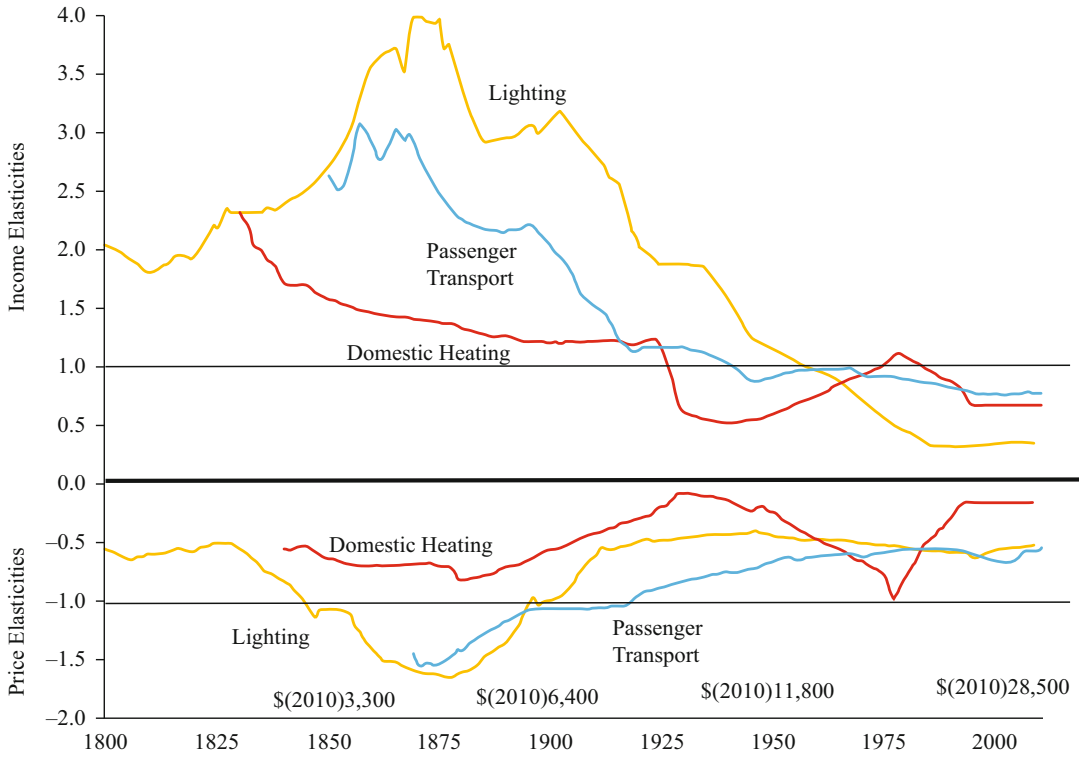
Energy service	Range of estimates	Number of studies
Space heating	−0.02 to −0.60	9
Space cooling	0.00 to −0.50	9
Water heating	−0.10 to −0.40	5
Lighting	−0.05 to −0.12	4
Transport (car)	−0.05 to −0.87	20

Source: Greening et al. (2000), Sorrell (2007), Sorrell and Dimitropoulos (2007), Azevedo (2014), Gillingham (2014) and Gillingham et al. (2016b)

Energy Services, Table 2 Price elasticities of United Kingdom demand for energy services used in scenarios towards a low carbon pathway

Residential sector services	Estimates	'Service' sector services	Estimates	Transport services	Estimates
Electrical appliances	−0.31	Electrical appliances	−0.32	Car	−0.54
Gas appliances	−0.33	Cooking	−0.23	Bus	−0.38
Space heating	−0.34	Space heating	−0.26	Rail (passenger)	−0.24
Water heating	−0.34	Water heating	−0.26	Rail (freight)	−0.24
		Lighting	−0.32	Goods Vehicles	−0.61
		Cooling	−0.32	Air travel	−0.38

Source: Adapted from Anandarajah et al. (2009)



Energy Services, Fig. 1 Income and price elasticities of demand for energy services in the United Kingdom, 1800–2010 (*Source*: Fouquet (2014))

realistic assumptions, including changes in energy service demand at different phases of economic development (as will be discussed in the next section), in long run scenarios of energy consumption and climate mitigation strategies, as prepared by the IPCC and the IEA.

Energy Services and Economic Development

Energy services have been increasingly linked to the debate about the role of energy access for economic and sustainable development. Energy services are seen as key to stimulating economic growth and development, and to ensuring improving living standards (Modi et al. 2006; AGECC 2010; UNDP 2011).

This then feeds through into greater consumption of energy services. Davis et al. (2014), in a rigorous analysis of Mexican households, find evidence of increased electricity consumption associated with air conditioning following improvements in the efficiency of equipment – in other words, there appear to be very large rebound effects. Sorrell (2007), for instance, argued that the direct rebound effect in developing countries may be larger since the demand for energy services may be far from saturated. In general, the hypothesis is, and the limited evidence suggests, that price and income elasticities of demand for energy services are greater in developing economies.

The main finding from Fouquet (2014) is that, as the United Kingdom's economy developed over the last 200 years, trends in income elasticities followed an inverse U-shape curve (see Fig. 1, top half). For instance, they reached a peak (about 2.3, 3.0 and 4.0 for income elasticities of demand for heating, transport and lighting, respectively) in the nineteenth century (at levels of GDP per capita below $(2010) 6,000). After the peaks, there were,

at first, rapid declines, then more gradual declines. Income elasticities took almost 100 years to reach unity (that is, a 10% increase in income led to a 10% rise in energy service consumption), in the mid-twentieth century, at between $(2010) 9–12,000 per capita. The results also indicate that income elasticities were significantly different from zero at high levels of per capita income in the twenty-first century, implying that current increases in income generate rises in energy service consumption (roughly, around 5% rises for a 10% increase in income). These rises feed through directly into greater energy consumption.

Developing economies may well also experience inverse U-shaped income elasticities, given saturation effects – that is, an additional unit of energy service generates less benefit or utility to the consumer. However, whether they peak and reach unit elasticity at similar levels of per capita GDP as the United Kingdom is unclear. Because today's developing economies have access to cheaper energy services (compared with the United Kingdom at the same level of income), they may experience earlier peaks (Fouquet 2014; van Benthem 2015).

These results offer the beginnings of a stylised fact about the relationship between elasticities of demand and economic development (Fouquet 2008, 2014). Sovacool (2011) describes the process as the 'energy service ladder'. That is, at very low levels of economic development, consumers focus on meeting basic needs, particularly food and cooking. As income grows, shelter and indoor climate become important – such as space and water heating, in temperate climates. As income rises further, these demands start to grow less proportionately than income (for example, income and price elasticities for heating fall). In turn, other demands are met, for instance, mobility, lighting and entertainment (implying rising income and price elasticities for transport and lighting demand). As income increases further, these income and price elasticities start to fall. Thus, pending confirmation from further studies, these general patterns could help to guide forecasts of energy service and, therefore, energy consumption. For example, while the IEA (2014) does incorporate saturation into its models (thus implying declining elasticity through time), they do not take account of the likelihood of peaking elasticities in developing economies.

As mentioned at the beginning of this section, while rising income drives up demand, rising consumption of energy services is likely to stimulate economic and social development – although it is hard to disentangle the direction of causality. This is made even harder by the idea that energy services can affect economic development through a number of different channels, and that these effects can change at different levels of economic development (Toman and Jemelkova 2003). Access to modern sources of energy for heating, cooking and power can bring about substantial health benefits, associated with reducing exposure to indoor air pollution or providing clean water and refrigeration, which can in turn yield improvements in productivity. Equally, they can enable a reallocation of household time (particularly for women) which can stimulate additional livelihood opportunities and improved education. Lighting may allow for greater flexibility in time allocation through the day and evening, as well as better conditions for education. Finally, lower transportation and communication costs may enable greater market size and access. In other words, although it can sometimes be hard to identify in the macroeconomic data, there appears to be a close relationship between electricity access and economic development (Toman and Jemelkova 2003; Modi et al. 2006; AGECC 2010; UNDP 2011).

Fuel poverty in general, and especially in developing economies, has major social consequences. 'Worldwide, approximately 3 billion people rely on traditional biomass for cooking and heating, and about 1.5 billion have no access to electricity. Up to a billion more have access only to unreliable electricity networks. The "energy-poor" suffer the health consequences of inefficient combustion of solid fuels in inadequately ventilated buildings, as well as the economic consequences of insufficient power for productive income-generating activities and for other basic services such as health and education. In particular, women and girls in the developing world are disproportionately affected in this regard' AGECC (2010, p. 7).

However, these health, education and welfare benefits tend to be ignored by policy-makers in developing economies (Reddy et al. 2009). Looking at experiences in Brazil, Bangladesh and South Africa, Winkler et al. (2011) stress that, despite access, affordability limits the ability to meet demands for specific energy services, and that policies addressing affordability appear to have more success in stimulating low energy-intensive services, such as lighting and entertainment, than high-intensive ones, such as cooking and cooling. Reddy (2015) discusses practical ways to make available affordable and reliable energy service to poor and often rural populations. One recommendation is to promote the development of small enterprises to provide relatively basic energy technologies. However, the implementation and scaling-up of the provision of energy supplies to meet service demands will need the close collaboration among numerous different stakeholders including households, local bodies, energy utilities, governments, entrepreneurs, research organisations, non-governmental organisations, community groups, financial institutions, and international agencies. Inevitably, coordination failures are a major barrier to enabling these multiple stakeholders to achieve the objectives in socially desirable ways.

Sovacool (2011) highlights how thinking about energy services emphasizes the role culture and social values play in influencing energy use. Indeed, the challenges of governing the development and expansion of energy markets will differ in each country partly because of the cultural aspects. For instance, an awareness of the value of travelling long distances to eat turkey with relatives in late November in the US or the value placed on well-ironed clothes on Sunday mornings in Uganda inform us about national patterns of energy service demands.

Providing an in-depth study of energy service behavior in Mexico, Cravioto et al. (2014) confirm that services are prioritised differently as incomes rise. Furthermore, they stress that the ability to measure the levels of satisfaction or utility generated may be easier by focusing on energy services. With this in mind, they find high levels of utility associated energy services provided to poor populations. However, they find that there is a relatively rapid declining marginal utility as energy service uses and incomes rise.

Concluding Discussion

This piece has sought to introduce the reader to the concept of energy services. This piece has shown why it is important to take account of energy service demand. Ignoring services, when analyzing energy markets, (especially when looking at the long run, where technical efficiencies of appliances and equipment can change considerably) is likely to lead to mis-estimation of price trends, mis-specfication of models, and biases in estimates.

The debate about the rebound effect, and identifying the actual energy savings resulting from efficiency improvements, has created a major increase in the interest in energy services. These empirical studies have shown that the non-zero price elasticity of demand implies that, after improving technical efficiency, consumers increase their consumption of energy services, but also generally reduce their energy consumption - though not by the same percentage as the efficiency improvements due to generally small, but non-negligible, rebound effects.

One of the historical barriers to using energy services in energy economics was that it 'distanced' the analysis from the influence of energy producers and suppliers. Particularly following the 1970s oil crises, and the growing role of OPEC, energy economics had tended to focus on energy supply and market structures (Fouquet 2013). Since the 1990s, environmental concerns have driven energy economists' research agendas, and issues related to the demand have become more important. This has meant a growing interest in incorporating energy end-use and service consumption.

As mentioned before, another limitation of this approach to understanding energy consumption behaviour (and a barrier to becoming the dominant modelling method) is the lack of data. Information about aggregate production and consumption by broad fuel categories is readily

available. Detailed data on end-use energy consumption, on energy efficiency or on energy services require far more effort and expense for statistical agencies.

A conclusion of this paper is, therefore, that there is a need to coordinate the methodological development for the collection of data on energy end-use and energy services consumption and prices across national statistical agencies, and encourage the collection of this data. Once this data becomes readily available, over time and across countries and regions, energy economists will be able to model and analyze the drivers of energy demand more accurately. This is likely to improve the reliability of future energy consumption and carbon dioxide emission scenarios. Furthermore, this information will enable stakeholders to observe the success of policies aimed at providing cheaper energy services while reducing energy use.

With this in mind, another important recommendation is that governments should be developing policies that seek the decoupling of energy services from energy (Fouquet 2015). They ought to create packages of measures, including targeted energy efficiency investments, that encourage more service consumption (which is welfare-enhancing), and less energy use and carbon emission (which is welfare-reducing). In other words, they need to develop policies that focus explicitly on energy services. For instance, there is a long run trade-off between lower energy prices and higher investment in energy efficiency (Newell et al. 1999; Popp 2002). Here, it is proposed that governments should take account of the trade-offs between energy prices and efficiency investment in the long run and ideally find the optimal trade-off between them. Indeed, energy service policies should go beyond simply looking at balancing energy prices and technical efficiency. They should seek to integrate policies related to the pricing and provision of energy sources with those focusing on promoting energy efficiency improvements, including research, development and demonstration (R,D&D) and considerations about behavioural features to address the energy efficiency gap (Gillingham and Palmer 2014) – and not exclusively through efficiency standards, which have received considerable criticism (Anderson et al. 2011). Finally, the active development of energy service policies should seek a broader and more strategic approach to thermal comfort, mobility, illumination, entertainment and computing.

The need to integrate policies related to energy services is particularly important for developing economies. Indeed, Fouquet (2016) stresses that policies promoting cheap energy (through large energy infrastructure projects and fuel subsidies) tend to discourage energy efficiency investment and lock economies into energy-intensive consumption patterns for decades. In turn, this behaviour leaves these economies vulnerable to energy price shocks, inflation, trade balance deficits, political pressures from energy companies and environmental pollution. Thus, successful long run economic development depends partly on careful and balanced policies related to energy services.

Despite the statistical and institutional barriers, it is hoped that there is sufficient grounds to convince analysts and policy-makers of the value of focusing on energy services in analyzing energy markets and in formulating climate policy. For analysts, their models and data ought to be based on energy services. Policy-makers need to, first, set up the framework for collecting data on energy services, combining information about energy price and consumption with the technical efficiency of equipment, then use models and analysis to determine the appropriate strategies. This may help formulate policies that are more effective at achieving their economic, social and environmental objectives.

See Also

- ▶ Energy-GDP Relationship
- ▶ Energy Transitions
- ▶ Rebound Effects

Acknowledgments I would like to thank Mona Chitnis and Ken Gillingham for their valuable comments on this paper. Naturally, the usual disclaimer applies. Support for this research from the ESRC is gratefully acknowledged.

Bibliography

AGECC (Advisory Group on Energy and Climate Change). 2010. Energy for a sustainable future: report and recommendations, the Secretary-General's advisory group on energy and climate change. United Nations, New York, US. http://www.un.org/wcm/webdav/site/climatechange/shared/Documents/AGECC%20summary%20report%5B1%5D.pdf

Anandarajah, G., and N. Strachan. 2010. Interactions and implications of renewable and climate change policy on UK energy scenarios. *Energy Policy* 38(11): 6724–6735.

Anandarajah, G., N. Strachan, P. Ekins, R. Kannan, and N. Hughes. 2009. Pathways to a low carbon economy: Energy systems modelling. UKERC Energy 2050 Research Report 1, UKERC. www.ukerc.ac.uk

Anderson, S.T., I.W.H. Parry, J.M. Sallee, and C. Fischer. 2011. Automobile fuel economy standards: Impacts, efficiency and alternatives. *Review of Environmental Economics and Policy* 5(1): 89–108.

Ayres, R., and B. Warr. 2009. *The economic growth engine: How energy and work drive material prosperity*. Cheltenham/Northampton: Edward Elgar Publishing.

Azevedo, I.L. 2014. Consumer end-use energy efficiency and rebound effects. *Annual Review of Environment and Resources* 39: 393–418.

Becker, G.S. 1965. A theory of the allocation of time. *The Economic Journal* 75: 493–517.

Beenstock, M., and P. Willcocks. 1981. Energy consumption and economic activity in industrialized countries: The dynamic aggregate time series relationship. *Energy Economics* 3(4): 225–232.

Berndt, E.R., and D.O. Wood. 1975. Technology, prices and the derived demand for energy. *Review of Economics and Statistics* 57(3): 259–268.

Boardman, B. 2010. *Fixing fuel poverty: Challenges and solutions*. London/New York: Earthscan.

Chitnis, M., and S. Sorrell. 2015. Living up to expectations: Estimating direct and indirect rebound effects for UK households. *Energy Economics* 52: S100–S116.

Cravioto, J., E. Yamasue, H. Okumura, and K.N. Ishihara. 2014. Energy service satisfaction in two Mexican communities: A study on demographic, household, equipment and energy related predictors. *Energy Policy* 73(C): 110–126.

Davis, L. 2008. Durable goods and residential demand for energy and water: Evidence from a field trial. *RAND Journal of Economics* 39(2): 530–546.

Davis, L., and P. Gertler. 2015. Contribution of air conditioning adoption to future energy use under global warming. *Proceedings of the National Academy of Sciences* 112(19): 5962–5967.

Dimitropoulos, J., L.C. Hunt, and G. Judge. 2005. Estimating underlying energy demand trends using UK annual data. *Applied Economics Letters* 12(4): 239–244.

Dubin, Jeffrey A., and Daniel L. McFadden. 1984. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52(2): 345–362.

Dubin, J.A., A.K. Miedema, and R.V. Chandran. 1986. Price effects of energy-efficient technologies: A study of residential demand for heating and cooling. *RAND Journal of Economics* 17(3): 310–325.

Fouquet, R. 2008. *Heat, power and light: Revolutions in energy services*. Cheltenham/Northampton: Edward Elgar Publications.

Fouquet, R. 2011. Divergences in long run trends in the prices of energy and energy services. *Review of Environmental Economics and Policy* 5(2): 196–218.

Fouquet, R. 2013. Introduction. In *Handbook on energy and climate change*, ed. R. Fouquet. Cheltenham/Northampton: Edward Elgar Publications.

Fouquet, R. 2014. Long run demand for energy services: Income and price elasticites over 200 years. *Review of Environmental Economics and Policy* 8(2): 186–207.

Fouquet, R. 2015. Lessons from energy history for climate policy. GRI Working Paper 209. Grantham Research Institute on Climate Change and the Environment, London School of Economics, London, UK.

Fouquet, R. 2016. Path dependence in energy systems and economic development. *Nature Energy*.

Fouquet, R., and P.J.G. Pearson. 2012. The long run demand for lighting: Elasticities and rebound effects in different phases of economic development. *Economics of Energy and Environmental Policy* 1(1): 83–100.

Frondel, M., J. Peters, and C. Vance. 2008. Identifying the rebound: Evidence from a German household panel. *The Energy Journal* 29(4): 145–163.

Fujimori, S., M. Kainuma, T. Masui, T. Hasegawa, and H. Dai. 2014. The effectiveness of energy service demand reduction: A scenario analysis of global climate change mitigation. *Energy Policy* 75: 379–391.

Gillingham, K. 2014. Rebound effects. In *New Palgrave dictionary of economics*, ed. Steven Durlauf and Lawrence Blume. London: Palgrave Macmillan Publishing.

Gillingham, K., and N.W. Chan. 2015. The microeconomic theory of the rebound effect and its welfare implications. *Journal of the Association of Environmental & Resource Economists* 2(1): 133–159.

Gillingham, K., and K. Palmer. 2014. Bridging the energy efficiency gap: Policy insights from economic theory and empirical analysis. *Review of Environmental Economics & Policy* 8(1): 18–38.

Gillingham, K., H. Deng, R. Wiser, N. Darghouth, G. Nemet, G. Barbose, V. Rai, and C. Dong. 2016a. Deconstructing solar photovoltaic pricing: The role of market structure, technology, and policy. *The Energy Journal* 37(3): 231–250.

Gillingham, K., D. Rapson, and G. Wagner. 2016b. The rebound effect and energy efficiency policy. *Review of Environmental Economics and Policy* 10(1): 68–88.

Goldemberg, J., T.B. Johansson, A.K.N. Reddy, and R.H. Williams. 1985. An end use oriented energy strategy. *Annual Review of Energy and the Environment* 10: 613–688.

Greening, L.A., D.L. Greene, and C. Difiglio. 2000. Energy efficiency and consumption – The rebound effect – A survey. *Energy Policy* 28(6–7): 389–401.

Groscurth, H.-M., T. Bruckner, and R. Kümmel. 1995. Modeling of energy-services supply systems. *Energy* 20(9): 941–958.

Haas, R., and P. Biermayr. 2000. The Rebound effect for space heating: Empirical evidence from Austria. *Energy Policy* 28: 403–410.

Haas, R., and L. Schipper. 1998. Residential energy demand in OECD countries and the role of irreversible efficiency improvements. *Energy Economics* 20(4): 421–442.

Haas, R., N. Nakicenovic, and A. Ajanovic. 2008. Towards sustainability of energy systems: A primer on how to apply the concept of energy services to identify necessary trends and policies. *Energy Policy* 36(11): 4012–4021.

Hausman, Jerry A. 1979. Individual discount rates and the purchase and utilization of energy-using durables. *Bell Journal of Economics* 10(1): 33–54.

Houthakker, H.S. 1951. Some calculations on electricity consumption in Great Britain. *Journal of the Royal Statistical Society. Series A* 114(3): 359–371.

Howarth, R.B. 1997. Energy efficiency and economic growth. *Contemporary Economic Policy* 15(4): 1–9.

Hunt, L.C., and D.L. Ryan. 2015. Economic modelling of energy services: Rectifying misspecified energy demand functions. *Energy Economics* 50: 273–285.

IEA. 2014. *World energy model documentation*. Paris: International Energy Agency, OECD.

Jevons, W.S. 1865. *The coal question: An inquiry concerning the progress of the nation, and the probable exhaustion of our coal-mines*. London: Macmillan Publishers.

Khazzoom, J.D. 1980. Economic implications of mandated efficiency standard for household appliances. *The Energy Journal* 1(4): 21–40.

Klein, Y.L. 1988. An econometric model of the joint production and consumption of residential space heat. *Southern Economic Journal* 55(2): 351–359.

Madureira, N.L. 2012. The anxiety of abundance: William Stanley Jevons and coal scarcity in the nineteenth century. *Environment and History* 18(3): 395–421.

Modi, V., S. McDade, D. Lallement, and J. Saghir. 2006. Energy and the Millennium Development Goals. Energy Sector Management Assistance Programme, United Nations Development Programme, UN Millennium Project, and World Bank, New York, US.

Muller, N.Z. 2016. On the divergence between fuel and service prices: The importance of technological change and diffusion in an American frontier economy. *Explorations in Economic History* 60: 93–111.

Neels, Kevin. 1981. Production functions for housing services. *Papers in Regional Science* 48(1): 25–37.

Newell, R.G., A.B.. Jaffe, and R.N. Stavins. 1999. The induced innovation hypothesis and energy-saving technological change. *Quarterly Journal of Economics* 114(3): 941–975.

Nordhaus, W.D. 1997. Do real output and real wage measures capture reality? The history of lighting suggests not. In *The economics of new goods*, ed. T.F. Breshnahan and R. Gordon. Chicago: Chicago University Press.

Olerup, B. 1998. Energy services a smoke screen. *Energy Policy* 26(9): 715–724.

Pearson, P.J.G. 2016. Energy transitions. In *New Palgrave dictionary of economics*, ed. Steven Durlauf and Lawrence Blume. London: Palgrave Macmillian Publishing.

Pindyck, R.S. 1979. Interfuel substitution and the industrial demand for energy: An international comparison. *Review of Economics and Statistics* 61(2): 169–179.

Popp, D. 2002. Induced Innovation and Energy Prices. *American Economic Review* 92(1): 160–180.

Quigley, J.M. 1984. The production of housing services and the derived demand for residential energy. *Rand Journal of Economics* 15(4): 555–567.

Quigley, J.M., and D. Rubinfeld. 1989. Unobservables in consumer choice: Residential energy and the demand for comfort. *Review of Economics and Statistics* 71(3): 416–425.

Reddy, B.S. 2015. Access to modern energy services: An economic and policy framework. *Renewable and Sustainable Energy Reviews* 47: 198–212.

Reddy, B.S., P. Balachandra, H. Salk, and K. Nathan. 2009. Universalization of access to modern energy services in Indian households – Economic and policy analysis. *Energy Policy* 37(11): 4645–4657.

Reister, D.B., and W.D. Devine 1979. Total costs of energy services ORAU/IEA-79-I?(R). Institute for Energy Analysis, Oak Ridge Associated Universities, Oak Ridge, US.

Reister, D.B., and W.D. Devine. 1981. Total costs of energy services. *Energy* 6(4): 305–315.

Römer, B., P. Reichhart, J. Kranz, and A. Picot. 2012. The role of smart metering and decentralized electricity storage for smart grids: The importance of positive externalities. *Energy Policy* 50: 486–495.

Saunders, H.D. 1992. The Khazzoom–Brookes postulate and neoclassical growth. *The Energy Journal* 13(4): 131–145.

Schleich, J., B. Mills, and E. Dütschke. 2014. A brighter future? Quantifying the rebound effect in energy efficient lighting. *Energy Policy* 72: 35–42.

Small, K.A., and K. van Dender. 2007. Fuel efficiency and motor vehicle travel: The declining rebound effect. *The Energy Journal* 28(1): 25–52.

Smulders, S., and M. de Nooij. 2003. The impact of energy conservation on technology and economic growth. *Resource and Energy Economics* 25(1): 59–79.

Sorrell, S. 2007. *The rebound effect: An assessment of the evidence for economy-wide energy savings from improved energy efficiency*. London: UK Energy Research Centre.

Sorrell, S. 2009. Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy* 37: 1456–1469.

Sorrell, S., and J. Dimitropoulos. 2008. The rebound effect: Microeconomic definitions, limitations and extensions. *Ecological Economics* 65(3): 636–649.
Sovacool, B.K. 2011. Conceptualizing urban household energy use: Climbing the "energy services ladder". *Energy Policy* 39(3): 1659–1668.
Stapleton, L., S. Sorrell, and T. Schwanen. 2016. Estimating direct rebound effects for personal automotive travel in Great Britain. *Energy Economics* 54: 313–325.
Strbac, G. 2008. Demand side management: Benefits and challenges. *Energy Policy* 36(12): 4419–4426.
Taylor, L.D. 1975. The demand for electricity: A survey. *The Bell Journal of Economics* 6(1): 74–110.
Toman, M.A., and B. Jemelkova. 2003. Energy and economic development: An assessment of the state of knowledge. *The Energy Journal* 24(2): 93–112.
Turner, K. 2013. Rebound' effects from increased energy efficiency: A time to pause and reflect. *The Energy Journal* 34(2).
UNDP. 2011. Human Development Report. United Nations Development Programme, ongoing initiatives by government, civil societies, and private sector companies to promote modern energy services. United Nations Development Programme.
van Benthem, A. 2015. Energy leapfrogging. *Journal of the Association of Environmental and Resource Economists* 2(1): 93–132.
Walker, I.O., and F. Wirl. 1993. Irreversible price-induced efficiency improvements: Theory and empirical application to road transportation. *The Energy Journal* 14(4): 183–205.
Weiller, C.M., and M. Pollitt. 2013. Platform markets and energy services. Cambridge Working Papers in Economics. Faculty of Economics, University of Cambridge, Cambridge, UK.
Winkler, H., A.F. Simões, E. Lèbre la Rovere, M. Alam, A. Rahman, and S. Mwakasonda. 2011. Access and affordability of electricity in developing countries. *World Development* 39(6): 1037–1050.

Energy Transitions

Peter J. G. Pearson

Abstract

This article explains why past, present and future energy transitions matter and why there is so much current interest in them in both developing and industrialised countries. It explores past transitions, including the first and subsequent industrial revolutions – and shows that although energy transitions have proceeded at various speeds in different places and times, and some of the more recent transitions have been faster, transitions do not usually happen quickly. An examination of lock-in, path dependence and the role of incumbents explains why there can be considerable inertia in energy systems and their technologies and institutions – but also suggests that incumbents may have an important role to play in a low-carbon transition. A brief review of recent studies in the area of sustainability transitions shows how researchers have aimed to understand transitions as interacting, co-evolving socio-technical processes and how studies that were mainly qualitative have now achieved better integration with quantitative analyses. A concluding discussion explores the many challenges involved in the governance and implementation of modern purposeful transitions, particularly low-carbon transitions. In these transitions, private incentives to adopt low-carbon fuels, technologies and practices are currently insufficient to ensure their adoption to the extent that makes sense for society; at the same time, the urgency of addressing greenhouse gas emissions from fossil fuels calls for unprecedentedly rapid change and the use of well-targeted, sustained but flexible policies and instruments.

Keywords

Biomass fuel; Climate change; Energy policy; Energy service; Energy transition; Environment; Fossil fuel; Greenhouse gas; Industrial revolution; Nuclear power; Path dependence; Primary energy; Renewable energy; Socio-technical transition

JEL Classifications

Q32; Q43; Q57; O40

What are Energy Transitions and Why Do They Matter?

An energy transition is often simply described as a shift from one dominant energy carrier to another. In practice, energy transitions involve changes or

shifts in how, where and by whom energy is produced, converted, supplied and used. These changes lead to new patterns, quantities and qualities of fuels, technologies and uses that interact and co-evolve with wider socioeconomic, demographic, technological and environmental developments and patterns: consequently, energy transitions are now often known as socio-technical transitions.

Energy transitions have included shifts from the dependence of early humans on firewood and human labour, to the growing use of animal labour and more sophisticated processing and uses of biomass fuels (peat, wood, grass, crop residues, charcoal), to wind and water power, coal, oil, gas and electricity. Such transitions have unfolded over decades and even centuries and are ongoing (Fouquet 2010; Smil 1994, 2010). Although the new sources may grow and dominate, the incumbent energy source(s) and technologies often continue to be used for several decades or longer (Fouquet 2008; Kander et al. 2013, Ch. 5).

Energy transitions matter because they have enabled – and been influenced by – increases in economic growth, welfare and population and the exploitation of natural resources. Economic history has shown how they have contributed greatly to human welfare through enabling significant, sustained increases in productivity and economic output and the production of new commodities and services (Kander and Stern 2014; Mokyr 2009). For many developing and emerging nations, transitions that provide affordable access to modern fuels, energy technologies and end-uses to large and growing populations are crucial elements of their development strategies (Barnes et al. 2005; IEA 2014; 2015a, b). Transitions also matter because different transitions and fuel mixes lead to different profiles of resource use and depletion (for renewable and fossil fuels respectively). And because fuels have different chemical properties (e.g. different fossil fuels have different ratios of hydrogen to carbon), and footprints in extraction, capture and use, transitions lead to different spatial and temporal environmental implications in the form of short-or long-lived local, regional and global impacts on air, land and water.

Policy on current transitions includes a focus on moves towards low-carbon fuels and technologies to reduce greenhouse gas emissions, particularly those from fossil fuels, and so address the threat of climate change (IPCC 2014). In most past transitions, however, there were obvious benefits that producers and consumers could gain from switching to the new energy sources and their uses, so such transitions were largely endogenous. Such private benefits are as yet much less evident for low-carbon technologies and practices. Both this gulf between private and social benefits and the widely (although not universally) perceived urgency of addressing climate change (Capstick et al. 2015) mean that a low-carbon transition has to be purposefully guided through public policy, a challenging contrast with most previous transitions (Pearson and Foxon 2012). Moreover, the avoidance of damages from climate change constitutes a global public good, i.e. one that is non-rival (one nation's benefit from avoided emissions and concentrations does not reduce the benefit available to others) and non-excludable (nations cannot be excluded from the benefits of avoided damage, even if they have not contributed towards such avoidance); this implies that it needs to be provided and financed via a form of global governance that is proving hard to construct.

Studies of past energy transitions include: Smil (1994, 2000) at an international scale; a range of international studies in Fouquet and Pearson (2012a); Kander et al. (2013), on five centuries of European experience; Schurr and Netschert (1960) on energy in the USA from 1850s, and more recently O'Connor and Cleveland (2014) on US transitions since 1780; and for the UK, Fouquet and Pearson (1998), Warde (2007) on primary energy transitions since 1760, Fouquet (2008, 2014) on energy services and Arapostathis et al. (2013) on the gas industry. Reviews and commentaries on energy transitions include Araújo (2014), Elzen et al. (2004), Fouquet (2015), Grin et al. (2010), Grubler (2008, 2012), Markard et al. (2012) and Smil (2010).

Transitions occur in the use of forms of primary energy (energy embodied in sources which involve human induced extraction or capture, so

as to make the energy available for trade, use or transformation), such as coal, oil, sunlight and wind. Transitions also occur in secondary energy forms or carriers delivered to the final user, such as gasoline, electricity and hydrogen. They help produce valued energy services, such as heat and comfort, mobility and illumination, with the aid of delivery infrastructures (pipes and wires) and end-use devices, such as light bulbs or passenger cars (Fouquet 2008). These secondary forms of energy are often of higher quality, such that they can be used for a wider and/or more valuable range of economically productive or satisfying activities (Cleveland et al. 2000; Stern 2010; Gentilvaite et al. 2015). They often cost more because of the conversion processes and losses involved (e.g. electricity and gasoline cost more than the fuels used to obtain them). However, users have been willing to pay for them because of their greater value and range of application in particular uses. For example, electricity is more flexible in use and, with efficient electric motors, enabled higher productivity than mechanical power from coal; gaseous and liquid fuels were essential for the internal combustion engine and more efficient, more flexible transport. Consequently they and their associated end-use devices tend to be increasingly demanded when incomes and living standards rise (Fouquet 2008, 2015; Grubler 2008), as recent experience in India, China and elsewhere confirms.

Recent Growth of Interest in the Study of Energy Transitions

Although energy transitions have long been of interest to many disciplines, academic and public policy interest in energy transitions grew steeply from the 1970s, partly in response to the stimuli of higher oil prices and environmental concerns. While economic historians have long studied them, energy transitions were of much less importance to mainstream economics until the past half-century or so of growth in the sub-disciplines of energy and environmental and development economics.

The 1970s saw rising concerns over oil availability and prices, resource depletion and energy-related environmental pollution, especially in the industrialised world, following the 1973–74 and 1979–80 oil price shocks; and in the developing world over 'the other energy crisis' (Eckholm 1975), i.e. worries about shortages of biomass fuels (e.g. wood fuel, charcoal, crop residues and animal dung), postulated impacts of wood fuel collection on deforestation (contested, since deforestation has many other more significant causes, such as land clearance for agriculture) (Anderson 1987) and other forms of environmental degradation, including soil erosion and health impacts from indoor air pollution from unenclosed stoves (Barnes et al. 2005; Fullerton et al. 2008).

From the late 1980s, along with continuing debate about petroleum resource depletion, the volatile geopolitics of oil and gas and ideas of sustainable development, we have seen a tightening policy focus on transitions to low-carbon forms of energy (renewables and nuclear electricity) and behavioural changes towards more efficient and reduced use of energy.

This arose *inter alia* from studies of the global climate impacts of concentrations of greenhouse gases associated with emissions from fossil fuels and other energy-related human activities, including cement production and land use change (IPCC 2014), and growing emphasis on the bringing together of climate policy and energy policy (Pearson and Watson 2012). Recent developments in seismic and drilling technologies, along with hydraulic fracturing ('fracking'), have led to significant exploitation of shale and unconventional gas and oil resources, especially in the USA. This has also brought a new set of environmental implications, both negative and positive, as well as impacts beyond the USA on the relative price and availability of coal and its use in power generation, and in some jurisdictions issues about the public acceptability of fracking (Hammond et al. 2015; Joskow 2015; Rasch and Köhne 2015; Stevens 2013).

Today, there are serious concerns about various ongoing and prospective transitions and their economic, environmental and social implications. These concerns include the challenges of mitigating and adapting to the very long-run impacts of

potential climate change; disquiet over the implications of the very rapid growth in fossil fuel use and associated urban air pollution and greenhouse gas emissions in recently industrialising countries like China and India; and the desire in many developing countries for rapid transitions from traditional biomass to modern fuels. As noted, common to all these areas is that they imply active management and guiding of transitions, which was largely absent from earlier energy transitions, while low-carbon energy transitions require a new form of global energy governance. They also imply largely unprecedented moves away from highly valued energy-dense fossil fuels towards less energy-and power-dense (and in some cases intermittent) forms of renewable energy (such as wind), which bring their own challenges and opportunities (Smil 2010).

Past Transitions: The First and Subsequent Industrial Revolutions

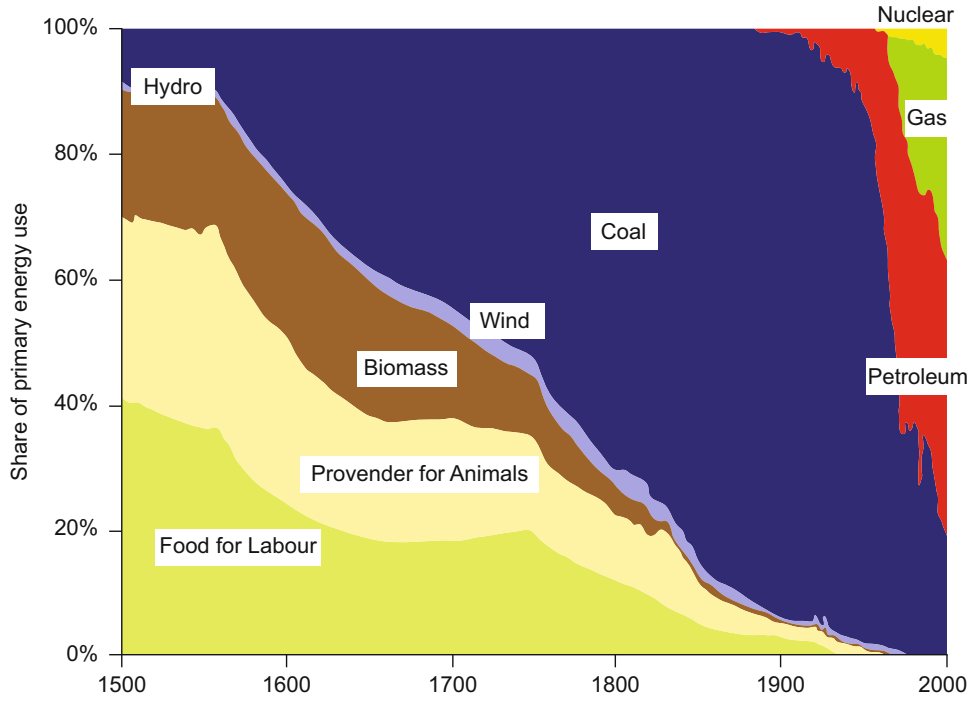
Although the past does not offer a blueprint for the future (Allen 2012), a knowledge of the characteristics, patterns and key relationships involved in past transitions can yield insights, parallels and partial analogues that may be instructive for policy thinking today. Indeed, Grubler (2012) suggests that part of the value of historic transitions work is that it helps develop 'storylines' for future transitions and invites us to question prevailing policy wisdom. It is not surprising, therefore, that part of the attention now paid to past transitions has been stimulated by the current concerns just described.

There has been rising interest in understanding how past transitions unfolded and in collating, recovering and analysing data on them (Fouquet 2008; Kander et al. 2013). Energy transitions have played major roles in industrial revolutions that have transformed economy and society. One of the most studied examples of a long, slow transition is that to coal before and during the first Industrial Revolution in Britain (in which much of the key activity took place in the 18th and 19th centuries). The second Industrial Revolution in the late 19th and early 20th centuries, which saw particularly striking developments in the USA and Germany, was intimately bound up with developments in petroleum, the internal combustion engine and electricity. Today, developments in the use of ICT in the energy system, including in 'smart grids', 'smart controls' and the 'internet of things', sometimes described as part of a third Industrial Revolution, could have major implications for energy use and economic growth and welfare; moreover, energy transitions in the developing world have the potential to transform – or not – the lives and life chances of billions whose living standards are constrained by lack of access to affordable modern fuels and technologies (GEA 2012a).

The long-drawn-out transition from wood and charcoal to coal in Britain before and during the first Industrial Revolution offers valuable insights into the complexities, causes and consequences of energy transitions, even though so much has changed since then. Figure 1 shows how the shares of different energy sources in Britain evolved over 500 years.

Wrigley (1988, 2010) shows how the 16th to 19th century transition from the limited energy flows possible in a constrained 'organic', largely biomass-based energy system to one based on coal helped transform Britain's economy in the Industrial Revolution. In the 'organic' economy – apart from intermittent wind and water power, energy *flows* were limited to what could be captured each year with available technologies and knowledge via photosynthesis. The organic material could feed people and draft animals and their labour and be used to provide heat and other energy services. The growing exploitation of *stocks* of coal, the fossilised, energy-dense accumulation from past photosynthesis, relaxed these constraints: 'To us the evidence points to the impossibility of sustaining high levels of growth or transformation in a world wholly dependent on 'organic' or vegetable sources of energy' (Kander et al. 2013, pp. 14–15).

Innovations, including the steam engine, the substitution of coal and coke for wood and charcoal in metal manufacture, new spinning and weaving technologies and textile mills, along with other socio-economic changes, helped lead

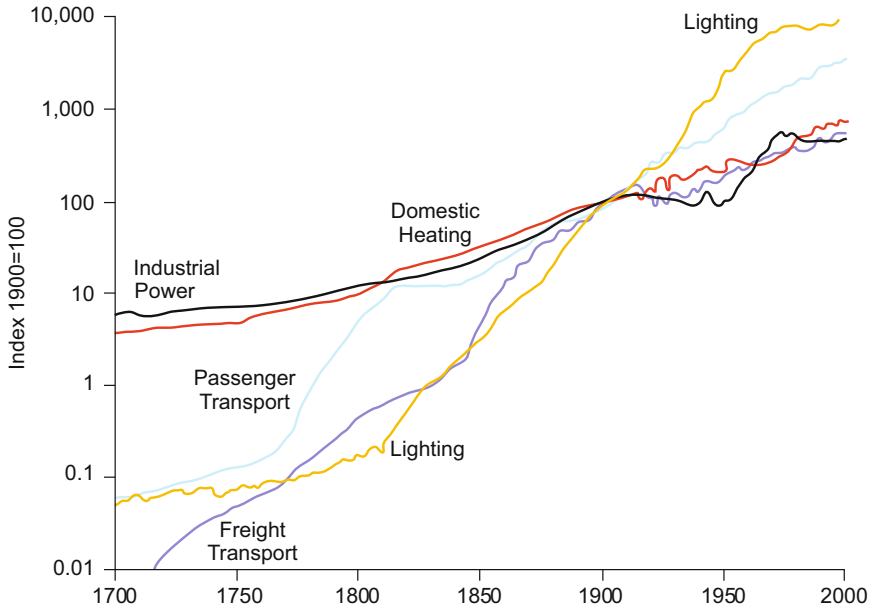


Energy Transitions, Fig. 1 Shares of primary energy consumption in Britain, 1500–2000 (Source: Reproduced from Fig. 5 in Fouquet (2010), with permission from the author and Elsevier. http://www.sciencedirect.com/science/article/pii/S0301421510004921. Original data from Fouquet (2008), with updates.)

and sustain the urbanisation, mechanisation and industrialisation of the Industrial Revolution. They led to striking declines in the cost of direct energy services (Fouquet 2008; Fouquet and Pearson 1998, 2006, 2012b; Pearson and Fouquet 2003) and increases in their consumption (see Fig. 2). They also led to efficiency improvements in manufacturing and other processes that yielded higher productivity and economic growth and better quality products. And they gradually meant much higher standards of living for the general population, as well as producing new problems of land, air and water pollution.

Unlike later transitions outside the UK, much of the transition to coal was pre-industrial; however, by around 1750, on the eve of the main years of the Industrial Revolution, coal already provided half of England's fuel (see Fig. 1). By 1900, 140 years later, almost all of the country's energy came from coal. Allen (2012, p. 17) states that 'Britain's transition to coal was bound up – both as cause and as effect – with the Industrial Revolution. . . High British wages and cheap coal underpinned the Industrial Revolution by creating a demand for technology that substituted capital and energy for labour. In Asia and much of Europe, low wages and dear energy had the opposite effect.'

The high wage/cheap energy price structure came from Britain's foreign trade boom in the 17th and 18th centuries. London's growing demand for fuel for industrial and domestic heating energy, a rising price of wood fuel relative to coal and the development and construction of the coal-burning house also created incentives to shift to coal and the means of doing so. On the supply side, Allen concentrates on the role of increased literacy and numeracy and the connections between the scientific discoveries of the 17th century and technological advances. For steam power he argues that the kind of R&D needed to bring the pan-European science to fruition, 'was more profitable in Britain than elsewhere, which is why the Industrial Revolution was invented in



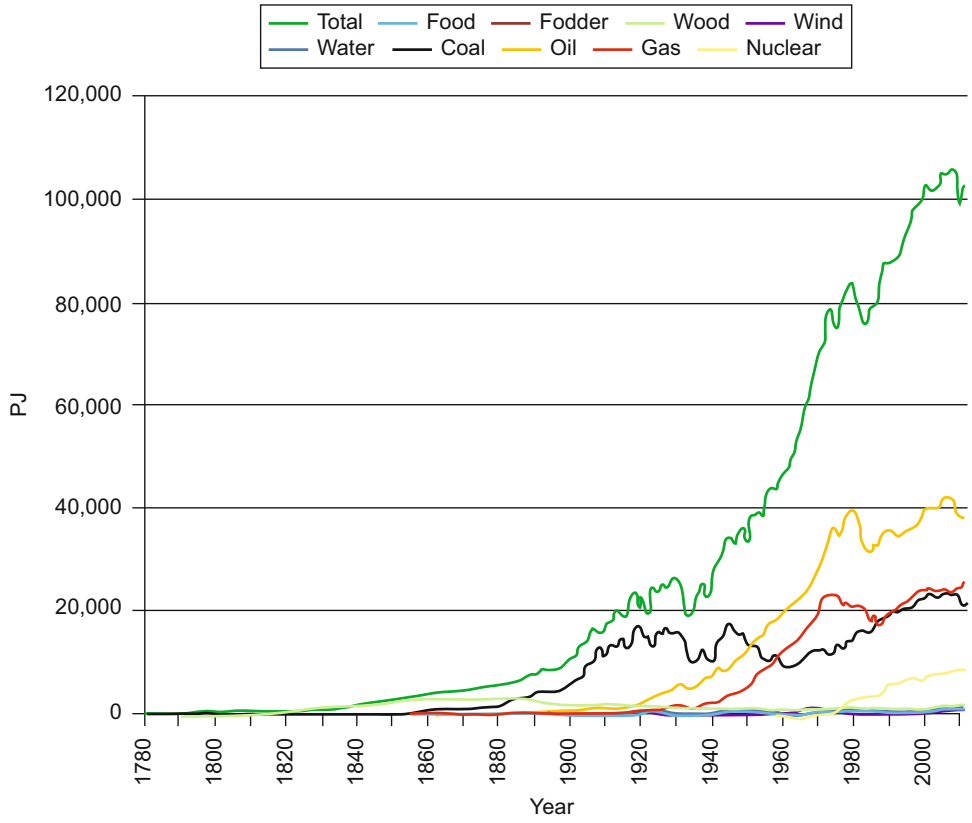
Energy Transitions, Fig. 2 Consumption of energy services in the UK, 1700–2000 (index 1900 = 100) (Source: Fig. 1 in Fouquet (2014). Reproduced with author's permission, under a Creative Commons license. http://creativecommons.org/licenses/by/4.0/.)

Britain.' The 'macro-invention' of the Newcomen engine was followed by a series of 'micro-inventions' that more than decimated the steam engine's coal consumption, to the point where the cost of coal became irrelevant to its commercial application and the steam engine became an 'appropriate technology' for other countries with different relative price structures.

Mokyr places much more emphasis than Allen on the view that the Industrial Revolution grew out of 'the social and intellectual foundations laid by the Enlightenment and the Scientific Revolution' (Mokyr 2009, p. 11). Britain led the Industrial Revolution because it could exploit its favourable human and physical resource endowment 'thanks to the great synergy of the Enlightenment: the combination of the Baconian program in useful knowledge and the recognition that better institutions created better incentives' (Mokyr 2009, 122). What was needed 'was the right combination of useful knowledge generated by scientists, engineers and inventors to be exploited by a supply of skilled craftsmen in an institutional environment that produced the correct incentives for entrepreneurs' (Mokyr 2009, p. 116).

The period between the late 1900s and the early 20th century, sometimes said to run between 1870 and 1914, with some precursor activity from the 1850s (although the boundaries are disputed), 'saw the lusty childhood, if not the birth... of a cluster of innovations that have earned the name of the Second Industrial Revolution' (Landes 1969, p. 235), which variously influenced the economies of the USA and Western Europe (especially Germany), with smaller and later effects in Eastern Europe and some parts of Asia and South America. These inventions and activities included many that were directly to do with energy, including electrical power, motors and lamps, telegraphy, the telephone and radio, the internal combustion engine and vehicles that used it and the growing use of oil and its products, as well as developments in organic chemicals and synthetics, in steel production, in the factory system and mass production and a range of advances in medical knowledge, public and private health and sanitation.

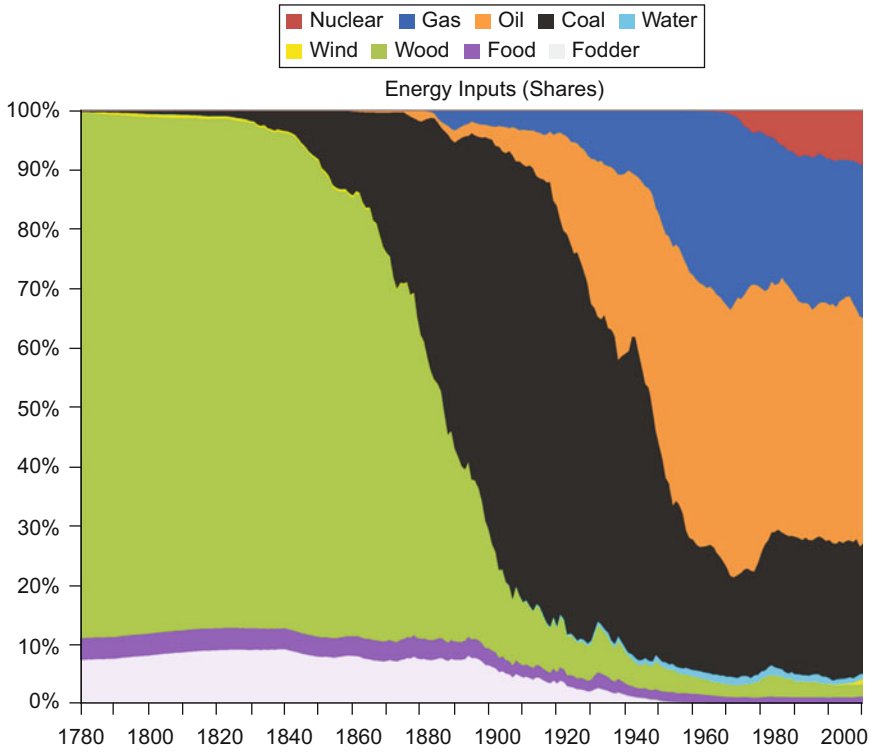
Mokyr (1999) argues that the path-breaking inventions of the second Industrial Revolution were crucial not because they necessarily had a



Energy Transitions, Fig. 3 US energy consumption 1780–2010 (in petajoules) (Source: O'Connor and Cleveland (2014, Fig. 13); reproduced under a Creative Commons license and with the author's permission. http://creativecommons.org/licenses/by/4.0/. See also O'Connor (2014).)

major direct impact on production but because they raised the effectiveness of research and development in microinventive activity and led to a range of further applications. While the first Industrial Revolution had a very limited scientific base, 'the persistence and acceleration of technological progress in the last third of the nineteenth century was due increasingly to the steady accumulation of useful knowledge' from science and the interplay with novel techniques and learning from experience. Along with this, changes in the organisation of production and the growing exploitation of economies of scale and scope, partly through mass production, meant that the second Industrial Revolution saw the rise of the large technological system, as with electrical power (Hughes 1983). Indeed, Gordon (2000, p. 1) argues that 'it was the Second Industrial Revolution, not the first, that created the golden age of productivity growth' that was to follow, with concomitant rises in income, wealth and living standards.

It should not be thought, however, that the pattern and duration of the energy transitions experienced in Britain during and after the Industrial Revolution would be simply replicated in other countries with different economies and resource endowments. For example, O'Connor and Cleveland (2014) undertook a comprehensive study of US energy transitions, through the assembly and use of a database of energy use for 1780–2010 that includes both traditional forms of energy (food, fodder, firewood, wind, water etc.) and the usual commercial forms (fossil fuels, nuclear and renewable electricity etc.). Figure 3 shows the remarkable rise in total energy use of

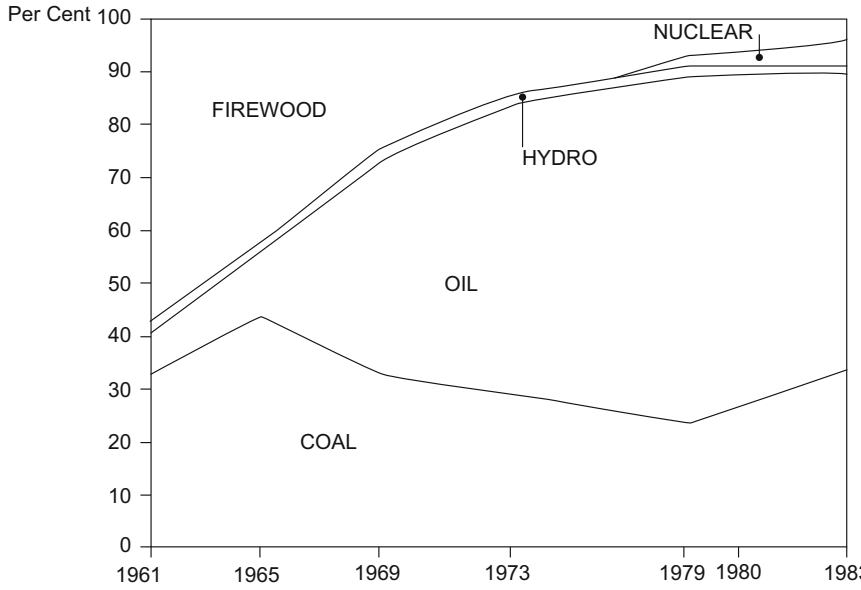


Energy Transitions, Fig. 4 Shares of inputs to US energy consumption, 1780–2010 (Source: O'Connor and Cleveland (2014, Fig. 14); reproduced under a Creative Commons license and with the author's permission. http://creativecommons.org/licenses/by/4.0/. See also O'Connor (2014).)

well over 300 times in 230 years (around 300 PJ (petajoules) in 1780 to 100,000 PJ by 2010). There is particularly rapid growth as the economy expanded in the 20th century, albeit with evidence of the impacts of major events, including the Great Depression, world wars and the oil price shocks of the 1970s.

O'Connor and Cleveland discuss major US transitions, building on the assumption that a reasonable benchmark for the start of a transition is when a new energy source captures 5 % of primary energy use (Smil 2010). Figure 4 shows that, in contrast with the dominance of coal early in the British Industrial Revolution (Fig. 1), as the US developed, wood retained a 50 % share of total energy use even through the mid-1880s. Although coal reached 5 % in the mid-1840s, it only gained a half share in the late 1880s, four decades later (through its use in steam engines, for transport and stationary power, in iron production and electricity generation); its share peaked at about 75 % three decades further on. Oil, first produced there in 1859, did not significantly replace coal until the early 20th century, when demand grew with the spread of the internal combustion engine and new and cheaper supplies of oil could satisfy it. Oil did not exceed coal's share until about 1950, peaking at just under 50 % in the late 1970s, whereas natural gas overtook coal's share by about 1960.

O'Connor and Cleveland also examine the changing energy intensity of the American economy, in terms of its energy/GDP (E/GDP) ratio. It is often thought that energy intensity rises with development, then falls in an inverted-U shape as an economy matures. When traditional energy sources are included, however, the US graph shows an overall declining trend, albeit with a 40-year period from around 1880 to 1920 when it was constant or rose. They ascribe the changing



Energy Transitions, Fig. 5 South Korea: shares in total primary energy supply, 1961–83 (Source: Pearson (1988) and the sources described there.)

slope of the E/GP ratio largely to the influence of three factors: (a) changes in energy quality (which they define as the marginal increase in GDP flowing from the use of one additional heat unit of a fuel), with higher quality fuels substituting for lower quality fuels over time, as in Fig. 3; (b) improvements in the efficiency of energy end-use, from incremental improvements in existing technologies and from the adoption of new technologies (which they illustrate via the example of changes in lighting technologies, but which have arisen in other main energy conversion processes and heat production); and (c) more efficient operational organisation and new manufacturing systems. And, of course, aggregate energy efficiency is influenced over time by the changing sectoral mix of activities with different energy intensities across the sectors of an economy (for a discussion of the reasons for the declining trend in aggregate energy intensity in Europe in the 20th century, see Gentilvaite et al. 2015).

Just as Britain's energy transitions were not replicated in nature or timing by those of the USA, Rubio and Folchi (2012) present evidence on the energy transitions from coal to oil for 20 Latin American countries over the first half of the 20th century. They argue that these small energy consumers had earlier and faster transitions than leading nations. By outlining a whole series of energy transitions, they identify a number of different transition processes. Factors such as domestic energy resources, the size of the internal market for energy services, trade relations and policy decisions were important in determining the nature and speed of the transitions experienced. They suggest that the lessons will be particularly relevant for understanding the way in which non-pioneering countries might adopt low-carbon energy sources and technologies.

The experience of the Republic of Korea (South Korea) since the beginning of the 1960s provides a striking example of more recent and much more rapid and highly directed transitions, illustrated in Fig. 5.

The figure shows that in less than 20 years, Korea transited from nearly 60 % dependence on wood fuel (and serious deforestation problems) to 90 % dependence on modern commercial fuels (Pearson 1988). Total energy supplies grew at an average of more than 8 % per year between 1962 and 1979, at a time when Korea experienced remarkably rapid economic growth and structural

change. In less than two decades, the country was transformed, via an outward-oriented industrialisation strategy, from a poor developing country into a semi-industrial middle income country, even though it was poorly endowed with energy resources or a mineral base for heavy industry (Kim 1983).

The first of three major transitions was from fuel wood to anthracite coal during the first five-year plan (1962–66); the second saw the growing replacement of coal by oil during the second plan (1967–71); and the third saw transition away from heavy dependence on oil towards a wider mix of fuels, after the first oil shock (1973–74), and more strikingly, after the second shock (1979–80). These were highly directed transitions in which, until the second oil shock, the main aim of energy policy was to ensure that energy did not become an obstacle to economic growth, with efforts directed towards supplying cheap and often subsidised energy to critical sectors (electrical power and industry). The increase in world oil prices and import costs and other inflationary pressures after the second shock led the government to re-evaluate the expansionist policies it had pursued until that point (Kim 1983).

Kander et al. (2013), in a major study of energy and economic development in Europe, have recently analysed the unfolding of energy transitions in several countries over five centuries. They do this through exploring first the pre-industrial economies and then the First, Second and Third Industrial Revolutions. In their analysis, they employ the term 'development blocks' 'to describe the series of systems of technology, infrastructure, energy sources and institutions by which economic growth proceeded'. Development blocks are constantly evolving systems centred on a generic technology (Dahmén 1988; Enflo et al. 2008). For instance, they say that it was the combination of coal, steam engine and iron, the raw material from which much of the new technology was made, that characterised and drove the first Industrial Revolution in Britain and then Western Europe. Kander et al. argue that although the process of growth is fairly continuous, it has been 'achieved through fundamentally discontinuous processes involving major structural shifts that take time to achieve' because of a significant lag between early inventions and the widespread adoption of a technology. They argue that transitions to higher quality energy carriers and rising thermal efficiencies in machines have led to improvements in economic energy efficiency, indicated by the ratio of GDP to Energy (GDP/E).

In addition, technological shifts associated with development blocks and industrial revolutions have produced structural shifts (changes in the relative importance of different activities) that have also significantly affected economic energy efficiency, making technologies more affordable, sometimes with concomitant *rebound* or *take-back* effects from increased demand (see Turner (2013) for a review of rebound concepts).

As with most analyses, Kander et al. (2013) emphasise and explore the interplay between energy, economic growth and population. As part of this, they examine the growth in total energy consumption (E) by decomposing it into the effects of three factors, energy intensity (E/GDP) – the inverse of economic energy efficiency, per capita income (GDP/P) and population (P) (these latter comprise the *scale effect*):

$$E = (E/GDP) \times (GDP/P) \times P$$

They show, for example, that for Western Europe between 1820 and 1910, while energy grew on average at the rate of 2.04 % per year, population grew at 0.76 %, income per capita by 1.08 % and energy intensity by 0.19 %.

Later they explore the evolution of carbon emissions (C) and the key factors that have influenced it by decomposing the widely used Kaya Identity (Ogawa 1991) into carbon intensity (C/GDP), energy intensity, per capita income and population, each of which can exert a significant influence:

$$C = (C/E) \times (E/GDP) \times (GDP/P) \times P$$

For other examples of such decompositions, see Pearson and Fouquet (1996, 2006).

Although coal and steam power are widely acknowledged to be key elements in the first

Industrial Revolution, it has not proved easy to demonstrate formally their rapid impact on output and productivity growth (Crafts 2004). It should be noted, however, that in a broader European context Kander et al. (2013, p. 368; also Appendix A) argue strongly that 'energy is more important to economic growth than generally believed among economists'. They suggest that such economists' focus on overall efficiency gains may miss the 'capital deepening' effect (linked to the growing use of machinery stimulated by cheap energy) that in their view has played an essential part in raising labour productivity (see also Kander and Stern 2014). They also argue that scepticism about energy's role in explaining economic growth is partly a reflection of economic growth models set up with low-or zero-cost shares for energy.

Energy Transitions and Time

As we have seen, economic historians have argued that the coming together of the elements that made up the transition from wood to coal in the UK not only had deep historical roots but also took a considerable time to develop. Pearson and Foxon (2012) note how evolutionary economists, drawing on some earlier economic ideas of Kondratiev and Schumpeter, identified five 'long waves' of economic development. In these waves, while the application of innovative technologies and processes, such as the steam engine, electrification and mass production, drove growth, the full societal benefits were only realised when wider institutions and practices had time to adapt to them. It is argued that structural crises of adjustment tend to arise in the face of the widespread introduction of radical new technologies because suitable new institutions and industrial structures have to be established to accommodate them (Freeman and Perez 1988; Freeman and Louça 2001).

Grubler (2012), considers the speed at which transitions have taken place. He notes that at the global level, characteristic 'change over times' (Marchetti and Nakicenovic 1979) in primary energy range from 80 years for the growth of oil/gas/electricity replacing steam power to 130 years for the growth of steam power displacing pre-industrial renewable sources. He also observes that while European late adopters of new technologies achieved faster transitions through profiting from learning externalities that reduced the costs of later adoption, early adopters faced the challenge of sunk costs associated with human, technological and infrastructural capital.

Wilson (2012) has also explored how and why energy supply and end-use technologies take time to mature. He studied processes of scaling-up, formative phases and learning in the historical diffusion of energy technologies from the early 1900s. He used logistic growth functions to help establish the time from initial commercialisation to market saturation. He concluded that: (1) increases in unit size come after an often prolonged experimentation with many smaller-scale units; and (2) that the peak growth phase of an industry can lag these increases in unit size by up to 20 years. Correspondingly, for a low-carbon transition, he suggested that it may be risky to use low-carbon technology policies that push for big jumps in unit size before a 'formative phase' of experimentation with smaller-scale units.

Fouquet (2010) examined past energy transitions and their drivers in the UK between 1500 and 1920, by end-use energy service (for heat, power, light and transport) and sector, to help find common features relevant for future transitions. He gives greater weight to post-Industrial Revolution transitions and reminds us of the dangers of aggregation: partly from a lack of detailed data, many historical studies have tended to analyse broad transitions within an economy. However, as he shows, a 'single' energy transition (say, from wood fuel to coal) may be composed of several different transitions, some running in parallel and others at very different times.

Fouquet identified the opportunities to produce cheaper and/or better energy services as the main economic drivers in the 14 cases he examined (and in most cases he also identified catalysts for faster adoption). The new services began in niche markets or market segments. Here, despite their often initial higher costs, users might be willing to pay for their extra service attributes (e.g. greater ease, cleanliness or flexibility in use, or faster

speed of transport). These new energy sources and energy-using technologies could be subsequently refined (e.g. leading to lower fuel costs or enhanced energy conversion efficiencies) until they could compete successfully across the market with the incumbent service provider (e.g. the switch from horse to rail transport or from gas to electric light). This process of refinement, performance or quality improvement and cost reduction meant that on average the whole innovation chain took more than 100 years and the diffusion phase nearly 50 years. Fouquet also suggests that, based on past experiences, a complete transition to a low-carbon economy could be very slow. This, he concludes, suggests that 'early action and favourable conditions may be warranted to steer any transition to a low-carbon economy'.

The studies reported in Fouquet (2008) found remarkable increases over time in the efficiency of converting energy flows into energy service flows, which were particularly striking in the case of lighting. Indeed, Grubler (2012) states that past energy transitions were essentially driven by technological and associated institutional and organisational transformations in energy end-use: 'transitions in energy services, in which new technological combinations enabled entirely new, or vastly improved traditional services, at greater energy efficiency and ever falling costs in a virtual, self-re-enforcing positive feedback loop drove associated transitions in energy supply systems'. While accepting that energy demand and supply systems co-evolve, and that there have been transformative changes in supply systems and technologies, he asserts that in the absence of energy service demand changes, we would not have seen the kinds of radical energy supply changes that have emerged so impressively from studies of the past. This implies the particular importance for policy of acting on the end-use and demand side, as well as the supply side. For example, it is clear that a successful low-carbon transition requires the decarbonisation of both heating and transport systems, which currently often depend on liquid and gaseous fossil fuels, with many consumers deeply attached to their current cars and heating systems; moreover, many current policy strategies are predicated on significant behavioural changes by energy consumers on an unprecedented scale.

Numerous studies suggest that although energy transitions have proceeded at different speeds in different places and times, and some of the more recent transitions have been faster, transitions do not usually happen quickly (Sovacool 2016). When they do, they seem likely to have built on a foundation of precursor activities in areas including infrastructure, institutions, technology, niche experimentation and an openness to change. For example, although the headline events of the UK's transition from town gas (produced from coal) to natural gas from the North Sea took 10 years between 1967 and 1977, the seeds were sown in the 1940s as the newly nationalised industry accepted the need to respond to the challenge of rising costs and growing competition from coal, oil and electricity. It not only experimented with alternative technologies and feed stocks but also made major institutional and organisational changes that enabled it to respond rapidly to the discovery of natural gas in the North Sea and to decide to strand its hundreds of town gas-producing assets (Arapostathis et al. 2013).

Lock-in, Path Dependence and the Role of Incumbents

In energy transitions, the penetration of innovative new fuels and technologies depends on their ability to compete with and displace the incumbent fuels, technologies and their industries. This in turn is influenced by the strategies and reactions of those incumbents. Long-term technological systems change can be path-dependent, in that a system's future evolution depends on the past sequence of events that led to its current state, i.e. its 'evolution is governed by its own history' (David 2007). So a system state may be locked in because of particular historical experiences, creating barriers to moving to an alternative state, although the conditions leading to that lock-in may no longer obtain (David 2001; Foxon 2007).

Arthur (1994) showed that increasing returns relating to scale, learning, adaptation and network

effects can lead to technological lock-in, while North (1990) and Pierson (2000) suggested that increasing returns can also apply to institutions, including market or regulatory frameworks – such that rule systems become hard to alter, enabling incumbents to protect their interests (see also Foxon 2011). These processes have important implications for future low carbon transitions. Thus Unruh (2000, 2002) and Unruh and Carrillo-Hermosilla (2006) suggest that co-evolutionary processes and mutually reinforcing positive feedbacks have led to the lock-in of current fossil fuel energy systems, i.e. carbon lock-in, and hence to systemic barriers to investment in low-carbon technology systems that can retard low-carbon transitions.

Nevertheless, it has also been argued that if increasing returns to new alternatives can be set off, this may lead to virtuous cycles of rapid change. Thus Garud and Karnøe (2001) have suggested there can also be 'path-creation' by incumbents: entrepreneurs may choose to depart from the structures they have jointly created. Incumbents may also, of course, sometimes increase their competitiveness. The sailing ship effect or last gasp effect of obsolescent technologies (Rosenberg 1976) is postulated to occur where competition from new technologies stimulates improvements in incumbent technologies/firms (but see Mendonça 2013). Analyses of energy industries threatened by technological discontinuities have, for example, offered insights into how the UK gaslight industry eventually responded to the threat from newly introduced incandescent electric light (Arapostathis et al. 2013) and why incumbent automotive technologies might show a sudden performance leap (Furr and Snow 2014); and, on the other hand, how current analyses may overestimate new entrants' ability to disrupt incumbent firms and underestimate incumbents' capacities to see the potential of new technologies and to integrate them with existing capabilities (Bergek et al. 2013). Given the urgency of climate change and the time taken for new entrants to scale up their activities, growing attention is now being paid to the potential of encouraging large relatively high carbon incumbents' positive engagement with low-carbon transitions, as well as to limiting their ability to constrain them.

Sustainability Transitions Studies

The growing interest in sustainable energy futures (e.g. GEA 2012a,b) is reflected in recent theoretical and empirical work in areas ranging from innovation to low-carbon transition pathways and policies to achieve them. We briefly explore an approach, often called 'sustainability transitions', that has attracted recent academic and policy attention. In sustainability transition studies, researchers explore potential societal transformations in production and consumption that combine economic and social development with reduced pressures on the environment.

In their review, on which this section draws, Markard et al. (2012) explain how this new field of studies has grown, as those in the policy arena and social scientists have paid growing attention to how to promote and govern a low-carbon transition to sustainability. In transition studies, sectors like energy are viewed as socio-technical systems that comprise interacting networks of actors (people, firms etc.), broadly defined institutions, material artefacts and knowledge. Thus an energy transition might unfold over several decades, and involve many actors and lead to new products, services, user practices, business models and organisations, as well as changes in technological and institutional structures and impacts well beyond the energy sector. Sustainability transitions are, therefore, complex, long drawn-out processes, involving governance and guidance (Smith et al. 2005), through which systems shift towards more sustainable modes of production, consumption and living.

The idea of the socio-technical regime brings ideas from evolutionary economics together with insights from the history and sociology of technology. It emphasises how scientific knowledge, engineering practices and processes are socially embedded. The regime tends to persist unless destabilised, leading to the emergence of a new regime. Much of the interest in this area is in regime changes, transitions, and the factors that

might result in such destabilisations (e.g. Kemp et al. 1998). For example, Turnheim and Geels (2013), explored the factors that led to the decline and ultimate demise of the UK's coal mining industry from its late 19th century dominance of the world coal market.

Researchers in sustainability transitions have carried out numerous empirical studies of historical energy transitions. These studies drew on the notion of the multi-level perspective (MLP), an approach built on the researches of Kemp, Rip and Schot (Kemp et al. 2001; Rip and Kemp 1998). The MLP posited that transitions could arise from dynamic interactions between the three interconnected levels of niche, regime and landscape. Pressures on the regime from the landscape, such as higher world oil prices or concern over climate change, might help prise open windows of opportunity for innovations in niches (i.e. 'protected spaces' in which innovations can develop and that might be managed strategically); some innovations might then break through, leading perhaps to fundamental regime shifts (Geels 2002, 2005; Raven et al. 2015). Different interactions could then lead to several different types of transition pathway, including pathways to future energy systems (Geels and Schot 2007). This approach has been used to help develop forward-looking analyses of the challenge of developing low-carbon energy transition trajectories, as in Hammond and Pearson (2013) and Trutnevyte et al. (2014), which describe studies of pathways to the UK's legislated target of 80 % reductions in greenhouse gas emissions by 2050.

Notions of transition management brought transition research together with insights from the areas of complex systems theory and governance. From these ideas came procedures to try to guide ongoing transitions toward greater sustainability (Kemp and Loorbach 2006). The guiding principles for transition management were derived from thinking about existing sectors as complex, adaptive systems and understanding management as a reflexive and evolutionary governance process (Voss et al. 2009). Transition management has been attempted in the Netherlands but has proved challenging, while questions have been raised about the political feasibility of trying to 'manage' national level transitions through such processes.

Energy Innovation Systems approaches have constituted another key approach in transition studies. Truffer et al. (2012), review the energy-related areas of the emerging socio-technical innovation systems literature, which has aimed to address the interacting social and technical aspects of innovation processes. This literature ranges across four innovation system areas: those of national (NIS), regional (RIS), sectoral (SIS) and technological (TIS) innovation systems.

The NIS developed in the late 1980s, partly as a riposte to what was seen by its proponents as the failure of neo-classical economics to explain the major economic challenges of that period: 'The core assumption was that nationally specific institutional arrangements between science, policy and industry explained differences in innovation success among different countries (especially the technology leaders US, Germany and Japan)' (Truffer et al. 2012, p. 4). The other three innovation systems approaches started from a criticism of the original NIS for limiting systems within their national boundaries and ignoring wider influences and interactions, such as the role of multinational companies. However, of these areas, Truffer et al. argue that the TIS tradition has been much the most productive in the energy field (see also Markard et al. 2015). TIS has its roots in the seminal paper of Carlsson and Stankiewicz (1991), which drew on Dahmén's work on development blocks (Dahmén 1988; Enflo et al. 2008), mentioned earlier.

Studies have gone from looking at selected energy innovations in particular countries, often focusing on the functions of the innovation system necessary for it to operate well (Hekkert et al. 2007), to inter-country comparisons and to some regional and global analyses of technological innovation systems. Much of the focus is on Europe, but with growing attention to emerging economies. The work includes an analytical framework of an 'energy technology innovation system' that claims to produce innovation policy guidelines that 'diverge substantially from policies implied by partial perspectives on innovation' (Gallagher et al. 2012). Despite the

progress in energy innovation systems research, Truffer et al. (2012) rightly suggest that there is room for further integration of the four systems approaches, and for further conceptual and empirical developments, including in the analysis of longer term energy transitions and their dynamics.

While much of the research on energy transition pathways has been historical and qualitative, belated but increasing attention is being given to the development of forward-looking quantitative approaches (Li et al. 2015) and with their consistent integration with qualitative analyses (Trutnevyte et al. 2014; Turnheim et al. 2015). Such developments are essential if these approaches are to provide more effective insights into guidance and policy for low-carbon transitions.

Energy Transition Pathways, Scenarios and Policies

While most past energy transitions were not purposefully guided along particular trajectories or pathways, modern transitions are different. Thus, developing and emerging countries and their citizens seek rapid economic growth, poverty reduction and higher living standards through wider access to modern fuels and energy-using technologies (IEA 2014; 2015a, b). For many, except in niche applications, the most direct and cheapest route has been via the exploitation of fossil fuels and fossil-generated electricity, often – as in China and India – through particularly rapid growth in the use of carbon-intensive coal and oil (IEA 2015a, c; National Bureau of Statistics of China 2015; Central Statistics Office 2015). At the same time, as well as growing recognition of the need to tackle carbon emissions (e.g. Liu et al. 2013; Reddy 2014), there has been growing governmental concern and public disquiet over energy-related pollution, particularly over the cumulative health impacts of air pollution in cities.

At the global level, the perceived urgency of addressing climate change was reflected in the 2015 Paris Agreement under the auspices of the United Nations Framework Convention on Climate Change (IPCC 2014; UNFCCC 2015). Although the developing world has strong incentives to avoid the damaging impacts of climate change, to which many will struggle to adapt, there is tension between the urgency of development now and the mitigation of greenhouse gases (GHGs), evident in the demands for financial transfers and technology transfer from the richer countries. Few in the developing world need extra incentives to adopt new fuels and technologies, and might be pleased to leapfrog to the most modern technologies if they could access and afford them. But, as noted earlier, the private benefits of adopting currently more expensive low-carbon technologies and practices are much less than the societal benefits of doing so. This, along with the fact that many low-carbon technologies do not, as yet, possess evidently superior bundles of performance characteristics to the fuels that they replace, means that policymakers face an unprecedented challenge (Pearson and Foxon 2012).

The desire for directed transitions to modern and/or low-carbon energy has led to a proliferation of energy transition scenarios and pathways. Although many earlier low-carbon transition scenarios provided technological detail, they tended to over-rely on exogenous emission constraints and high-level trends, without paying sufficient attention to how policy, technology and behaviour might interact and how scenario trajectories and end-points might be achieved (Hughes and Strachan 2010). Recently, however, growing attention has been paid to scenario or pathway construction that, in line with some of the thinking in the sustainability transition studies area, acknowledges this interaction, incorporates the roles of different system actors and, rather than focusing on their notional endpoints, explores the means whereby the trajectories of transition pathways might be realised (e.g. GEA 2012a; Foxon 2013; Hammond and Pearson 2013; ETI 2015; RTP Engine Room 2015). Nevertheless, much remains to be done to make low-carbon scenarios and pathways contribute more effectively to our understanding of the challenges and dynamics of the transition and how to address them.

Any transition from fossil fuels poses significant challenges for electricity generation (widely

dependent on coal, natural gas and oil), for all forms of transport (widely dependent on petroleum-based fuels and natural gas), and for process heat and domestic and commercial heating, ventilation and cooling (HVAC) (widely dependent on coal, gas, oil and fossil-generated electricity). To respond to these challenges will require significantly more (low-carbon) electrified provision of transport, heat and HVAC, with corresponding implications for infrastructure investment and management, such as electric vehicle charging and hydrogen refuelling stations for fuel cell vehicles.

The decarbonisation of electricity, for example, raises technical, social, behavioural and financial issues. On the supply side, incorporating very large proportions of renewable energy brings issues of power density (the rate of flow of energy per unit of land area) (Smil 2010) and intermittent generation (e.g. from the variability of wind and sunlight). The latter requires compensating backup and/or storage capacity, while low utilisation of some of this capacity might undermine its efficiency and economic viability. Although any electrical system must balance supply and demand, greater use of renewables may place new short-run (e.g. hourly) and longer run (e.g. seasonal) challenges. These can be eased by better forms of storage and by managed demand side responses across the whole system to enhance the efficiency of generation, transmission and distribution network capacity and manage its operational and economic performance (Aunedi et al. 2013; Pudjianto et al. 2014).

While nuclear electricity is widely proposed as a potentially valuable element in a low-carbon portfolio, its technologies tend to be relatively capital-intensive (like many renewables), to be slow to construct and not infrequently to meet with public distrust about operational safety, waste disposal or proliferation risks. Consequently, in many countries it remains too risky for private investors, in the absence of the sustained comfort of long-term financial support from the state (Joskow and Parsons 2012). Its economic prospects may depend on the development of more modular, flexible, rapidly constructible, publicly acceptable designs and waste disposal strategies. Some countries, such as China, are proceeding with significant nuclear electricity programmes. In others, including Germany, however, nuclear power has met with considerable political opposition, exacerbated by concerns raised by the 2011 Fukushima Daiichi events.

Increasing interest has developed in various forms of decentralised or distributed electricity generation (and other forms of energy). As well as the technical challenges of interfacing them with larger transmission and distribution systems, they will also require regulatory, financial and business model innovation, which is likely to involve selling energy services, such as illumination and comfort (RTP Engine Room 2015). Existing electricity (and other energy) utilities face the challenge of adapting their business models, which have been predicated on the expansion of demand via large-scale, centralised technologies, and have often been associated with limited consumer satisfaction and trust (Richter 2013). Recent electricity utility reorganisations, such as those of the German-based companies E. ON in 2014 and RWE in 2015, indicate attempts to respond to some of these issues and opportunities (as well to as the 2011 government decision to phase out and close Germany's nuclear power plants by 2022).

Not least because of recent rapid and projected growth in coal-based generation in China (IEA 2015c) and India (IEA 2015a), emphasis has been placed on the development of systems of carbon capture and sequestration (CCS), e.g. in depleted petroleum reservoirs or salt caverns, at either pre-or post-combustion stages. Moreover, some low-carbon scenarios depend for their ultimate effectiveness on combinations of sustainable biomass with CCS that could yield negative carbon emissions. CCS faces issues that are technical, economic (imposition of a cost penalty), financial and social (in terms of public acceptability) (Hammond and Spargo 2014; Watson 2012). Similarly, ideas of geoengineering and climate manipulation, from solar radiation management to carbon dioxide removal, to offset the greenhouse effect or limit GHG concentrations, also raise analogous and perhaps more challenging

issues, as well as those of the international governance and stability of such a global undertaking (Bellamy and Lezaun 2015).

On the demand side, many scenarios or pathways incorporate significant restraint of the growth of energy demand. Such restraint comes partly from elements of behavioural change and acceptance of new 'smarter' technologies, ranging from monitoring and control of the use of domestic devices, like dishwashers, washing machines and refrigerators, to possible management and uses of the storage and discharge capacities of automotive batteries. A growing body of research suggests that while there is real interest in such developments, the modification of energy-using social practices is far from straightforward and will require much better understanding of different groups' knowledge, beliefs and habits if it is to occur. Long-run scenarios also tend to include significant and challenging modifications to and investment in the design, regulation and energy efficiency of the built environment, from dwellings to neighbourhoods and cities, including the retrofitting of existing structures (Dixon et al. 2014).

There has been some concern about possible rebound effects, i.e. whether, as energy services such as lighting, thermal comfort or transport become more efficiently delivered and hence cheaper, a significant amount of the efficiency is taken back in the form of increased consumption and energy use (Turner 2013). In this area, there are likely to be significant differences between situations of latent demand in rapidly growing emerging or developing countries, in which consumer demands may be highly responsive to increases in income and wealth and falling energy service prices, and mature economies where responsiveness may be expected to be much less (Fouquet 2015).

The Stern Review of the Economics of Climate Change (Stern 2007) argued that the innovation and deployment of low-carbon technologies requires at least three types of government policy measure: (1) a carbon price, through a carbon tax or tradable permit scheme; (2) direct support for research, development and demonstration (RD&D) and early stage commercialisation of low carbon technologies; and (3) measures to remove or address institutional and non-market barriers to the uptake of energy-efficient and low-carbon options.

In orthodox economic terms, these three measures address three types of 'market failure' that justify government intervention. The carbon price is there to 'internalise' the negative externality of climate change from CO_2. The innovation supports are there to harness the positive externalities or spillovers available from RD&D, including addressing the 'valley of death' between a technology's demonstration and the commercialisation phases. Following Kenneth Arrow, it is argued that the private market will underinvest in invention and research because such activities are risky, because social value exceeds private value and because of increasing returns in the use and scaling up of innovations. The institutional and non-market barriers are those that can delay or inhibit the commercialisation and growth of innovations; this also connects with ideas of path dependence and lock-in, mentioned earlier. Alternatively, and drawing on ideas from the innovation literature, it has also been argued that broader ideas of 'systems failures' should be seen as complements to or superior substitutes for the market failure approach (Bleda and del Rio 2013; Foxon 2015). This approach includes proposals that addressing large-scale societal challenges, including climate change, requires a long-term visionary and mission-oriented approach in which public investment should play a significant part (Mazzucato and Penna 2015).

There is continuing debate about the relative merits of carbon taxes (setting a 'price' to achieve a quantity reduction) and marketable permits (setting a quantity that achieves a price) as one of the routes to controlling the climate externality. And it is clear that there are significant practical challenges with both instruments, e.g. the problem of setting politically acceptable tax levels and the governance issues of agreeing and running a permit system across multiple jurisdictions and sectors (as in the EU tradable permits scheme). Moreover, distributional issues arise, since expenditures on energy normally form a larger proportion of lower than of higher incomes and there can

be serious concerns over levels of 'fuel poverty'. These issues can be addressed by the use of other policy instruments, such as income supports, provided that such instruments are available and workable.

Governments may provide financial and regulatory support for appropriate levels of low-carbon RD&D, via policy instruments such as renewable energy certificates (or 'green certificates'), feed-in tariffs (FiTs) and auctions for electricity generating capacity. This raises the question of how such supports should be paid for. They are sometimes paid for via energy consumers' bills, as in the UK. However, Newbery (2015) argues that the general principles of public finance imply that societal public goods like climate change mitigation should be financed from general taxation of the population at large. Several jurisdictions have recently shown greater interest in electricity generating capacity auctions, because a well-designed bidding process may reduce the cost of such supports. Moreover, if a support instrument is set up without appropriate exit strategies to phase it out when costs fall – because of learning, economies of scale or other sources of cost reduction – it may either prove hard to change or sudden changes may disrupt investor confidence.

Similarly, there are arguments for providing public funds for investment in low-carbon infrastructure, as now happens in countries like Germany, since the state may be able raise finance at lower cost than the private sector (Newbery 2015). Private sector investors may insist on risk premiums that reflect the perceived risks of new technologies, whose prospects depend on uncertain future state commitments to climate change targets and their associated carbon price trajectories. More broadly, a key challenge for the low-carbon transition is for the state, and ultimately the global community, to give sufficiently credible, consistent longer term commitments and to help create the policy structures and instruments to achieve them, while also retaining a necessary flexibility in the face of changing conditions.

Successful transitions to new fuels and technologies depend, as suggested earlier, not only on the cost and performance of low-carbon technologies, but also on what is happening to incumbent technologies, infrastructures and institutions. Fossil fuel incumbents use a variety of strategies to protect their interests (and in some cases even embrace the new technologies). Their position may be, however, considerably bolstered by the effective inverse of a carbon tax, namely a subsidy. Thus many countries maintain very large subsidies to fossil fuels and infrastructure investments, distorting relative prices and creating significant barriers to low-carbon fuels and developments. Estimates of such subsidies vary widely but are big. According to the International Energy Agency's (IEA) estimates, worldwide fossil fuel subsidies, measured as the gap between end-user prices and reference prices, totalled $490 billion in 2014; this is almost four times the size of subsidies to help deploy renewable energy technologies in the power sector of $112 billion, plus $23 billion for biofuels (IEA 2015a). Coady et al. (2015) offer a tentative but strikingly larger estimate, on the basis of post-tax subsidies, i.e. the subsidy that arises when consumer prices are below supply costs plus a tax to reflect environmental damage, including non-climate damage, and a tax applied to consumption goods to raise revenues; they estimate that in 2013 these subsidies amounted to $4.9 trillion (6.5 % of global GDP) and were continuing to rise.

The International Energy Agency (IEA) suggest in a 2015 *World Energy Outlook Special Report* (IEA 2015b) that several targeted policy actions might be done at zero net economic cost, leading to a peak in energy-related GHG emissions by 2020: increasing energy efficiency in the industry, buildings and transport sectors; progressively reducing the use of the least-efficient coal-fired power plants and forbidding their construction; raising investment in renewable energy technologies in the power sector from $270 billion in 2014 to $400 billion in 2030; phasing out remaining fossil fuel subsidies to end-users by 2030; and reducing methane emissions in oil and gas production (since methane is a potent GHG). Nevertheless, the achievement of such policies would require significant political will and public acceptance.

Since energy and climate policies and instruments are administered by regional, national and

local governments, governance matters. This is not only in terms of the efficiency and transparency of the processes of government and their implementation, but also in terms of the balance of policy objectives and the means whereby they are achieved. For example, for many countries policy has three overarching broad objectives, sometimes called the 'Energy Policy Trilemma'. They are: climate and environment; energy security; and affordability and cost. The balance of these objectives and the trade-offs between them tend to shift through both internal and external influences, as for example when energy security became a key objective in oil-importing countries after the oil price shocks of the 1970s. Such shifts can be a challenge for a stable policy trajectory towards a targeted energy transition.

Policy implementation can be variously in the hands of the state, the market and civil society, depending on the dominant governance framing or 'logic' of the time (Foxon 2013; Johnson et al. 2016). Frequently it involves a hybrid mix of two or even all three of these. In the UK, for instance, the dominant logic of the 1980s/1990s/early 2000s was that of the private market, reflecting the neoliberal ideology of Margaret Thatcher's governments and its influence on later administrations. After the Climate Change Act of 2008 committed successive UK governments to 80 % reductions in GHGs by 2050 from a 1990 base, and partly because of difficulties in meeting both climate change and energy security objectives purely via the private sector, the UK has moved towards a more hybrid market/state form of governance, along with growing interest in decentralised energy and greater involvement of civil society (Pearson and Watson 2012; Emamian 2014). While the preferred mix of logics varies by country and political system, it is clear that governments and citizens have had cause to reflect on how best to promote, govern and guide an unprecedented low-carbon transition.

See Also

- ▶ Climate Change, Economics of
- ▶ Energy Economics
- ▶ Energy Price Shocks
- ▶ Energy-GDP Relationship
- ▶ Industrial Revolution
- ▶ Oil and The Macroeconomy
- ▶ Rebound Effects

Acknowledgments Work on this article was supported by funding from the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant EP/K005316/1] under the 'Realising Transition Pathways' project. The author is solely responsible for the views expressed and any errors and omissions. He thanks Roger Fouquet and two reviewers for helpful and perceptive criticisms and suggestions.

Bibliography

Acemoglu, D., U. Akcigit, D. Hanley, and W. R. Kerr. 2014. *Transition to clean technology*. Harvard Business School Working Paper, No. 15–045. http://dash.harvard.edu/bitstream/handle/1/13506422/15-045.pdf?sequence=1

Allen, R. 2009. *The British industrial revolution in global perspective*. Cambridge: Cambridge University Press.

Allen, R. 2012. Backward into the future. *Energy Policy* 50: 17–23.

Anderson, D. 1987. *The economics of afforestation: A case study in Africa*. World Bank Occasional Paper No. 1/New Series. Johns Hopkins University Press, Baltimore.

Arapostathis, S., A. Carlsson-Hyslop, P.J.G. Pearson, J. Thornton, M. Gradillas, S. Laczay, and S. Wallis. 2013. Governing transitions: Cases and insights from two periods in the history of the UK gas industry. *Energy Policy* 62: 25–44.

Araújo, K. 2014. The emerging field of energy transitions: Progress, challenges and opportunities. *Energy Research & Social Science* 1: 112–121.

Arthur, W.B. 1994. *Increasing returns and path dependence in the economy*. Ann Arbor: University of Michigan Press.

Aunedi, M., P.A. Kountouriotis, J.E.O. Calderon, D. Angeli, and G. Strbac. 2013. Economic and environmental benefits of dynamic demand in providing frequency regulation. *IEEE Transactions on Smart Grid* 4(4): 2036–2048.

Barnes, D.F., K. Krutilla, and W.F. Hyde. 2005. *The urban household energy transition: Social and environmental impacts in the developing world*. Washington DC: Resources for the Future.

Bellamy, R. and, J. Lezaun. 2015. Crafting a public for geoengineering. *Public Understanding of Science*. 1–16. doi:10.1177/0963662515600965.

Bergek, A., C. Berggren, T. Magnusson, and M. Hobday. 2013. Technological discontinuities and the challenge for incumbent firms: Destruction, disruption or creative accumulation? *Research Policy* 42(6–7): 1210–1224.

Bleda, M., and P. del Rio. 2013. The market failure and the systemic failure rationales in technological innovation systems. *Research Policy* 42: 1039–1052.

Bruns, S.B., C. Gross, and D.I. Stern. 2014. Is there really granger causality between energy use and output? *Energy Journal* 35(4): 101–134.

Capstick, S., L. Whitmarsh, W. Poortinga, N. Pidgeon, and P. Upham. 2015. International trends in public perceptions of climate change over the past quarter century. *Wiley Interdisciplinary Reviews: Climate Change* 6(1): 35–61.

Carlsson, B., and R. Stankiewicz. 1991. On the nature, function and composition of technological systems. *Journal of Evolutionary Economics* 1(2): 93–118.

Central Statistics Office. 2015. Energy statistics 2015. Ministry of statistics and programme implementation, Government of India. New Delhi. http://mospi.nic.in/Mospi_New/upload/Energy_stats_2015_26mar15.pdf

Cleveland, C.J., R.K. Kaufmann, and D.I. Stern. 2000. Aggregation and the role of energy in the economy. *Ecological Economics* 32: 301–317.

Coady, D., I. Parry, L. Sears, and B. Shang. 2015. How large are global energy subsidies? IMF Working Paper WP/15/105. International Monetary Fund.

Crafts, N. 2004. Steam as a general purpose technology: A growth accounting perspective. *Economic Journal* 114: 338–351.

Crafts, N. 2010. Explaining the first industrial revolution: Two views. *European Review of Economic History* 15: 153–168.

Dahmén, E. 1988. 'Development blocks' in industrial economics. *Scandinavian Economic History Review* 36: 3–14.

David, P.A. 2001. Path dependence, its critics and the quest for 'historical economics'. In *Evolution and path dependence in economic ideas: Past and present*, ed. P. Garrouste and S. Ioannides, 15–40. Cheltenham: Edward Elgar.

David, P.A. 2007. Path dependence – A foundational concept for historical social science. *Cliometrica* 1(2): 91–114.

Dixon, T., M. Eames, M. Hunt, and S. Lannon. 2014. *Urban retrofitting for sustainability: Mapping the transition to 2050*. London: Routledge.

Eckholm, E. 1975. *The other energy crisis: Firewood*. Worldwatch Paper 1. Worldwatch Institute, Washington DC.

Elias, R. J., and D. G. Victor. 2005. *Energy transitions in developing countries: A review of concepts and literature*. Working Paper #40. Program on energy and sustainable development. Stanford University. http://www.trunity.net/files/158401_158500/158492/elias-and-victor-2005.pdf

Elzen, B., F.W. Geels, and K. Green. 2004. Conclusion. transitions to sustainability: Lessons learned and remaining challenges, Ch. 12. In *System innovation and the transition to sustainability: Theory, evidence and policy*, ed. B. Elzen, F.W. Geels, and K. Green. Cheltenham: Edward Elgar.

Emamian, S.M.S. 2014. British electricity policy in flux: Paradigm ambivalence and technological tension. PhD Thesis. University of Edinburgh.

Enflo, K., A. Kander, and L. Schön. 2008. Identifying development blocks – A new methodology. *Journal of Evolutionary Economics* 18: 57–76.

ETI. 2015 – Energy Technologies Institute. 2015. *Options, choices, actions: UK scenarios for a low carbon energy system transition*. Energy Technologies Institute, Loughborough. http://www.eti.co.uk/wp-content/uploads/2015/03/Options-Choices-Actions-Hyperlinked-version-for-digital.pdf

Fouquet, R. 2008. *Heat, power and light: Revolutions in energy services*. Cheltenham: Edward Elgar.

Fouquet, R. 2010. The slow search for solutions: Lessons from historical energy transitions by sector and service. *Energy Policy* 38(10): 6586–6596.

Fouquet, R. (ed.). 2013. *Handbook on energy and climate change*. Cheltenham/Northampton: Edward Elgar.

Fouquet, R. 2014. Long run demand for energy services: Income and price elasticities over 200 years. *Review of Environmental Economics and Policy* 8(2): 186–207.

Fouquet, R. 2015. *Lessons from energy history for climate policy*. Centre for climate change economics and policy Working Paper 235/Grantham Research Institute on climate change and the environment Working Paper 209. http://www.lse.ac.uk/GranthamInstitute/publication/lessons-from-energy-history-for-climate-policy/

Fouquet, R., and P.J.G. Pearson. 1998. A thousand years of energy use in the United Kingdom. *Energy Journal* 19(4): 1–41.

Fouquet, R., and P.J.G. Pearson. 2006. Seven centuries of energy services: The price and use of lighting in the United Kingdom (1300–2000). *The Energy Journal* 27: 139–177.

Fouquet, R., and P.J.G. Pearson. 2012a. Editorial: Past & prospective energy transitions: Insights from history. Introduction to special section. *Energy Policy* 50: 1–7.

Fouquet, R., and P.J.G. Pearson. 2012b. The long run demand for lighting: Elasticities and rebound effects in different phases of economic development. *Economics of Energy and Environmental Policy* 1: 83–100.

Foxon, T.J. 2007. Technological lock-in and the role of innovation, ch. 9. In *Handbook of sustainable development*, ed. G. Atkinson, S. Dietz, and E. Neumayer. Cheltenham/Northampton: Edward Elgar.

Foxon, T.J. 2011. A co-evolutionary framework for analysing transition pathways to a sustainable low carbon economy. *Ecological Economics* 70: 2258–2267.

Foxon, T.J. 2012. Transition pathways for a UK low carbon electricity future. *Energy Policy* 52: 10–24.

Foxon, T.J. 2013. Transition pathways for a UK low carbon electricity future. *Energy Policy* 52: 10–24.

Foxon, T.J. 2015. Rationale for policy interventions in sustainability transitions. Paper for 6th international sustainability transitions (IST) conference, SPRU, University of Sussex, Falmer.

Foxon, T.J., and P. Pearson. 2008. Overcoming barriers to innovation and diffusion of cleaner technologies: Some

features of a sustainable innovation policy regime. *Journal of Cleaner Production* 16(1, Supplement 1): S148–S161.

Freeman, C., and F. Louça. 2001. *As time goes by*. Oxford: Oxford University Press.

Freeman, C., and C. Perez. 1988. Structural crises of adjustment: Business cycles and investment behaviour. In *Technical change and economic theory*, ed. G. Dosi, C. Freeman, R. Nelson, G. Silverberg, and L. Soete. London: Pinter.

Fullerton, D.G., N. Bruce, and S.B. Gordon. 2008. Indoor air pollution from biomass fuel smoke is a major health concern in the developing world. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 102: 843–851.

Furr, N.R., and D.C. Snow. 2014. Intergenerational hybrids: Spillbacks, spillforwards, and adapting to technology discontinuities. *Organization Science* 26(2): 475–493.

Gallagher, K.S., A. Grübler, L. Kuhl, G. Nemet, and C. Wilson. 2012. The energy technology innovation system. *Annual Review of Environment and Resources* 37: 137–162.

Garud, R., and P. Karnøe. 2001. Path creation as a process of mindful deviation. In *Path dependence and creation*, ed. R. Garud and P. Karnøe. London: Lawrence Erlbaum.

GEA. 2012a. Global energy assessment: Toward a sustainable future, eds. T.B. Johansson, P. Anand, N. Nakicenovic, and L. Gomez-Echeverri. *International Institute for Applied Systems Analysis*. Cambridge University Press, Cambridge/Laxenburg. Also at: http://www.iiasa.ac.at/web/home/research/Flagship-Projects/Global-Energy-Assessment/Chapters_Home.en.html

GEA. 2012b. Global energy assessment: Toward a sustainable future (Key Findings, Summary for Policymakers, Technical Summary), eds. T.B. Johansson, P. Anand, N. Nakicenovic, and L. Gomez-Echeverri. *International Institute for Applied Systems Analysis*. Laxenberg. http://www.iiasa.ac.at/web/home/research/Flagship-Projects/Global-Energy-Assessment/GEA-Summary-web.pdf

Geels, F. 2002. Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. *Research Policy* 31: 1257–1274.

Geels, F.W. 2005. *Technological transitions and system innovations: A coevolutionary and socio-technical analysis*. Cheltenham: Edward Elgar.

Geels, F.W., and J. Schot. 2007. Typology of socio-technical transition pathways. *Research Policy* 36: 399–417.

Gentilvaite, R., A. Kander, and P. Warde. 2015. The role of energy quality in shaping long-term energy intensity in Europe. *Energies* 8: 133–153.

Gordon, R.J. 2000. Does the 'New Economy' measure up to the great inventions of the past? Working Paper 7833, *National Bureau of Economic Research*. Cambridge, MA. http://www.nber.org/papers/w7833

Grin, J., J. Rotmans, J. Schot, F. Geels, and D. Loorbach. 2010. *Transitions to sustainable development: New directions in the study of long term transformative change*. London/New York: Routledge.

Grubler, A. 2008. Energy transitions, *The encyclopedia of earth* [online]. http://www.eoearth.org/view/article/152561/

Grubler, A. 2012. Energy transitions research: Insights and cautionary tales. *Energy Policy* 50: 8–16.

Hammond, G.P., and P.J.G. Pearson. 2013. Challenges of the transition to a low carbon, more electric future: From here to 2050. *Energy Policy* 52: 1–9.

Hammond, G.P., and J. Spargo. 2014. The prospects for coal-fired power plants with carbon capture and storage: A UK perspective. *Energy Conversion and Management* 86: 476–489.

Hammond, G.P., Á. O'Grady, and D.E. Packham. 2015. Energy technology assessment of shale gas 'fracking' – A UK perspective. *Energy Procedia* 75: 2764–2771.

Hekkert, M., R.A.A. Suurs, S. Negro, S. Kuhlmann, and R. Smits. 2007. Functions of Innovation systems: A new approach for analysing technological change. *Technological Forecasting and Social Change* 74: 413–432.

Hughes, T.P. 1983. *Networks of power: Electrification in Western society, 1880–1930*. Baltimore: Johns Hopkins Press.

Hughes, N., and N. Strachan. 2010. Methodological review of UK and international low carbon scenarios. *Energy Policy* 38: 6056–6065.

IEA (International Energy Agency). 2014. Africa energy outlook. OECD/IEA, Paris. https://www.iea.org/publications/freepublications/publication/WEO2014_AfricaEnergyOutlook.pdf

IEA (International Energy Agency). 2015a. *India energy outlook*. World energy outlook Special Report. Paris: OECD/IEA. http://www.worldenergyoutlook.org/india/

IEA (International Energy Agency). 2015b. *Southeast Asia Energy outlook*. World energy outlook Special Report. Paris: OECD/IEA. http://www.iea.org/publications/freepublications/publication/WEO2015_SouthEastAsia.pdf

IEA (International Energy Agency). 2015c. *Special data release with revisions for people's Republic of China*. Paris: OECD/IEA. https://www.iea.org/publications/freepublications/publication/SpecialdatareleasewithrevisionsforPeoplesRepublicofChina04.11.2015.pdf

IPCC (Intergovernmental Panel on Climate Change). 2014. *Climate change 2014: Synthesis report*. Geneva: IPCC. https://www.ipcc.ch/pdf/assessment-report/ar5/syr/AR5_SYR_FINAL_SPM.pdf

Jacobsson, S., and V. Lauber. 2006. The politics and policy of energy system transformation – Explaining the German diffusion of renewable energy technology. *Energy Policy* 34: 256–276.

Johnson, V.C.A., F. Sherry-Brennan, and P.J.G. Pearson. 2016. Alternative liquid fuels in the UK in the inter-war

period (1918–1938): Insights from a failed energy transition. *Environmental Innovation and Societal Transitions*. doi:10.1016/j.eist.2015.12.001.

Joskow, P. 2015. The shale gas revolution: Introduction. *Economics of Energy and Environmental Policy* 4(1): 1–5.

Joskow, P.L. and, J.E. Parsons. 2012. *The future of nuclear power after Fukushima*. MIT Center for energy and environmental policy research Working Paper CEEPR WP2012-001. http://web.mit.edu/ceepr/www/publications/workingpapers/2012-001.pdf

Kander, A., and D. Stern. 2014. Economic growth and the transition from traditional to modern energy in Sweden. *Energy Economics* 46: 56–65.

Kander, A., P. Malanima, and P. Warde. 2013. *Power to the people: Energy in Europe over the last five centuries*. Princeton/Oxford: Princeton University Press.

Kemp, R., and D. Loorbach. 2006. Transition management: A reflexive governance approach, ch. 5. In *Reflexive governance for sustainable development*, ed. J.-P. Voss, D. Bauknecht, and R. Kemp. Cheltenham/Northampton: Edward Elgar.

Kemp, R., J. Schot, and R. Hoogma. 1998. Regime shifts to sustainability through processes of niche formation. The approach of strategic niche management. *Technology Analysis and Strategic Management* 10(2): 175–195.

Kemp, R., A. Rip, and J. Schot. 2001. Constructing transition paths through the management of niches. In *Path dependence and creation*, ed. R. Garud and P. Karnoe, 269–299. London: Lawrence Erlbaum.

Kim, Y.H. 1983. Rational and effective use of energy in Korea's industrialisation. *Energy* 8(1–2): 107–123.

Landes, D.S. 1969. *The unbound prometheus: Technological change and industrial development in Western Europe from 1750 to the present*. Cambridge: Cambridge University Press.

Li, F.G.N., E. Trutnevyte, and N. Strachan. 2015. A review of socio-technical energy transition (STET) models. *Technological Forecasting & Social Change* 100: 290, online.

Liu, Z., D. Guan, D. Crawford-Brown, Q. Zhang, K. He, and J. Liu. 2013. A low-carbon road map for China. *Nature* 500: 143–145.

Marchetti, C., and, N. Nakićenović. 1979. The dynamics of energy systems and the logistic substitution model. (Research Report 79–13). Laxenburg: International Institute for Applied Systems Analysis.

Markard, J., R. Raven, and B. Truffer. 2012. Sustainability transitions: An emerging field of research and its prospects. *Research Policy* 41: 955–967.

Markard, J., M. Hekkert, and S. Jacobsson. 2015. The technological innovation systems framework: Response to six criticisms. *Environmental Innovation and Societal Transitions* 16: 76–86.

Mazzucato, M., and C. Penna. 2015. *Mission-oriented finance for innovation: New ideas for investment-led growth*. London: Policy Network and Rowman and Littlefield International. http://www.policy-network.net/publications/4860/Mission-Oriented-Finance-for-Innovation

Mendonça, S. 2013. The 'sailing ship effect': Reassessing history as a source of insight on technical change. *Research Policy* 42: 1724–1738.

Mokyr, J. 1999. The second industrial revolution, 1870–1914. In *Storia Dell'economia Mondiale,* ed. V. Castronovo. Rome: Laterza. http://faculty.wcas.northwestern.edu/~jmokyr/castronovo.pdf

Mokyr, J. 2009. *The enlightened economy*. London: Penguin Books.

National Bureau of Statistics of China. 2015. *China statistical yearbook 2014*. China Statistics Press. http://www.stats.gov.cn/tjsj/ndsj/2014/indexeh.htm

Newbery, D. 2015. *Reforming UK energy policy to live within its means*. Energy Policy Research Group Working Paper 1516. Cambridge: University of Cambridge.

North, D.C. 1990. *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.

O'Connor, P.A. 2014. Aspects of energy transitions: History and determinants. Doctoral Dissertation. Boston University.

O'Connor, P.A., and C.J. Cleveland. 2014. U.S. energy transitions 1780–2010. *Energies* 7(12): 7955–7993.

Ogawa, Y. 1991. Economic activity and the greenhouse effect. *Energy Journal* 12(1): 23–35.

Pearson, P. 1988. *Energy transitions in less-developed countries: Analytical frameworks for practical understanding*. Energy Discussion Paper 40. Cambridge: Cambridge University Energy Research Group.

Pearson, P.J.G., and R. Fouquet. 1996. Energy efficiency, economic efficiency and future CO_2 emissions from the developing world. *The Energy Journal* 17(4): 135–160.

Pearson, P.J.G., and R. Fouquet. 2003. Long run carbon dioxide emissions and Kuznets curves: Pathways to development. In *Energy in a competitive market: Essays in honour of colin robinson*, ed. L.C. Hunt. Cheltenham: Edward Elgar.

Pearson, P.J.G., and R. Fouquet. 2006. Long run carbon dioxide emissions and Kuznets curves: Pathways to development. In *Energy in a competitive market*, ed. L. Hunt. Cheltenham: Edward Elgar.

Pearson, P.J.G., and T.J. Foxon. 2012. A low-carbon industrial revolution? Insights and challenges from past technological and economic transformations. *Energy Policy* 50: 117–127.

Pearson, P.J.G. and, J. Watson. 2012. *UK energy policy 1980–2010: A history and lessons to be learned*. London: IET/The Parliamentary Group for Energy Studies. http://www.theiet.org/factfiles/energy/uk-energy-policy-page.cfm?type=pdf

Pearson, P.J.G. and, J. Watson. 2013. *UK energy policy 1980–2010: A history and lessons to be learnt: A review to mark 30 years of the parliamentary group for energy studies*. IET and Parliamentary Group for Energy Studies. http://www.theiet.org/factfiles/energy/uk-energy-policy-page.cfm?type=pdf

Pierson, P. 2000. Increasing returns, path dependence, and the study of politics. *American Political Science Review* 94(2): 251–267.

Pudjianto, D., M. Aunedi, P. Djapic, and G. Strbac. 2014. Whole-systems assessment of the value of energy storage in low-carbon electricity systems. *IEEE Transactions on Smart Grid* 5(2): 1098–1109.

Rasch, E.D., and M. Köhne. 2015. Hydraulic fracturing, energy transition and political engagement in the Netherlands: The energetics of citizenship. *Energy Research & Social Science*. doi:10.1016/j.erss.2015.12.014.

Raven, R., F. Kern, A. Smith, S. Jacobsson, and B. Verhees. 2015. The politics of innovation spaces for low-carbon energy. Introduction to the special issue. *Environmental Innovation and Societal Transitions*. doi:10.1016/j.eist.2015.06.008.

Reddy, B.S. 2014. *India's energy transition – Pathways for low-carbon economy.* WP-2014-025. Mumbai: Indira Gandhi Institute of Development Research. http://www.igidr.ac.in/pdf/publication/WP-2014-025.pdf

Richter, M. 2013. Business model innovation for sustainable energy: German utilities and renewable energy. *Energy Policy* 62: 1226–1237.

Rip, A., and R. Kemp. 1998. Technological change. In *Human choice and climate change – Resources and technology*, ed. S. Rayner and E.L. Malone, 327–399. Columbus: Battelle Press.

Rosenberg, N. 1976. *Perspectives on technology.* Cambridge: Cambridge University Press.

RTP Engine Room – Realising Transition Pathways Engine Room. 2015. *Distributing power: A transition to a civic energy future*. Realising Transition Pathways Research Consortium. Bath: University of Bath. http://www.realisingtransitionpathways.org.uk/realisingtransitionpathways/news/FINAL_distributing_power_report_WEB.pdf

Rubio, M.S., and M. Folchi. 2012. Will small energy consumers be faster in transition? Evidence from the early shift from coal to oil in Latin America. *Energy Policy* 50: 50–61.

Schurr, S.H., and B.C. Netschert. 1960. *Energy in the American economy, 1850–1975. An economic study of its history and prospects*. (Resources for the Future, Inc.). Baltimore: Johns Hopkins University Press.

Smil, V. 1994. *Energy in world history.* Boulder/London: Westview Press.

Smil, V. 2000. Energy in the twentieth century: Resources, conversions, costs, uses, and consequences. *Annual Review of Energy & Environment* 25: 21–51.

Smil, V. 2010. *Energy transitions: History, requirements, Prospects*. Santa Barbara: Praeger.

Smith, A., A. Stirling, and F. Berkhout. 2005. The governance of sustainable socio-technical transitions. *Research Policy* 34: 1491–1510.

Sovacool, B.K. 2016 (in press). How long will it take? Conceptualizing the temporal dynamics of energy transitions. *Energy Research and Social Science*. http://www.sciencedirect.com/science/article/pii/S2214629615300827

Stern, N. 2007. *The economics of climate change: The stern review*. Cambridge: Cambridge University Press.

Stern, D.I. 2010. Energy quality. *Ecological Economics* 69(7): 1471–1478.

Stevens, P. 2013. Gas markets: Past, present and future, Ch. 5. In *Handbook on energy and climate change*, ed. R. Fouquet. Cheltenham/Northampton: Edward Elgar.

Truffer, B., Markard, J., Binz, C., and S. Jacobsson. 2012. *A literature review on energy innovation systems*. EIS Radar paper [online]. http://www.eis-all.dk/~/media/Sites/EIS_Energy_Innovation_Systems/english/about%20eis/Publications/EIS_Radarpaper_final.ashx

Trutnevyte, E., J. Barton, Á. O'Grady, D. Ogunkunle, D. Pudjianto, and E. Robertson. 2014. Linking a storyline with multiple models: A cross-scale study of the UK power system transition. *Technological Forecasting and Social Change* 89: 26–42.

Turner, K. 2013. 'Rebound' effects from increased energy efficiency: A time to pause and reflect. *Energy Journal* 34(4): 25–42.

Turnheim, B., and F.W. Geels. 2013. The destabilisation of existing regimes: Confronting a multi-dimensional framework with a case study of the British coal industry (1913–1967). *Research Policy* 42(10): 1749–1767.

Turnheim, B., F. Berkhout, F.W. Geels, A. Hof, A. McMeekin, B. Nykvist, and D.P. van Vuuren. 2015. Evaluating the sustainability transitions pathways: Bridging analytical approaches to address governance challenges. *Global Environmental Change* 35: 239–253.

UNFCCC (United Nations Framework Convention on Climate Change). 2015. *Adoption of the paris agreement*. FCCC/CP/2015/L.9/Rev.1. Paris UNFCCC. http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf

Unruh, G.C. 2000. Understanding carbon lock-in. *Energy Policy* 28: 817–830.

Unruh, G.C. 2002. Escaping carbon lock-in. *Energy Policy* 30: 317–325.

Unruh, G.C., and J. Carrillo-Hermosilla. 2006. Globalizing carbon lock-in. *Energy Policy* 34: 1185–1197.

Van Den Bergh, J.C.J.M., B. Truffer, and G. Kallis. 2011. Environmental innovation and societal transitions: Introduction and overview. *Environmental Innovation and Societal Transitions* 1(1): 1–23.

van der Ploeg, R. 2011. Macroeconomics of sustainability transitions: Second-best climate policy, Green Paradox, and renewable subsidies. *Environmental Innovation and Societal Transitions* 1(1): 130–134.

Voss, J.-P., A. Smith, and J. Grin. 2009. Designing long-term policy: Rethinking transition management. *Policy Sciences* 42: 275–302.

Warde, P. 2007. *Energy consumption in England and wales 1560–2000*. Napoli: ISSM-CNR.

Watson, J. ed. et al. 2012. Carbon capture and storage: Realising the potential? UKERC Report UKERC/RR/ESY/CCS/2012/001. London: UK Energy Research Centre. http://www.ukerc.ac.uk/support/tiki-download_file.php?fileId=2421

Wilson, C. 2012. Up-scaling, formative phases, and learning in the historical diffusion of energy technologies. *Energy Policy* 50: 81–94.
Wrigley, E.A. 1988. *Continuity, chance and change. The character of the industrial revolution in England*. Cambridge: Cambridge University Press.
Wrigley, E.A. 2010. *Energy and the english industrial revolution*. Cambridge: Cambridge University Press.

Energy-GDP Relationship

David I. Stern

Abstract

Energy use and GDP are positively correlated, although energy intensity has declined over time and is usually lower in richer countries. Numerous factors affect the energy intensity of economies, and energy efficiency is obviously one of the most important. However, the rebound effect might limit the possibilities for energy efficiency improvements to reduce energy intensity. Natural science suggests that energy is crucial to economic production but mainstream economic growth theory largely ignores the role of energy. Ecological economists and some economic historians argue that increasing energy supply has been a principal driver of growth. It is possible that historically energy scarcity imposed constraints on growth, but with the increased availability of modern energy sources energy's importance as a driver of growth has declined. Empirical research on whether energy causes growth or vice versa is inconclusive, but meta-analysis finds that the role of energy prices is central to understanding the relationship.

Keywords

Ecological economics; Economic growth; Energy; Energy intensity; GDP; Granger causality

JEL Classifications

Q32; Q43; Q57; O40

Introduction

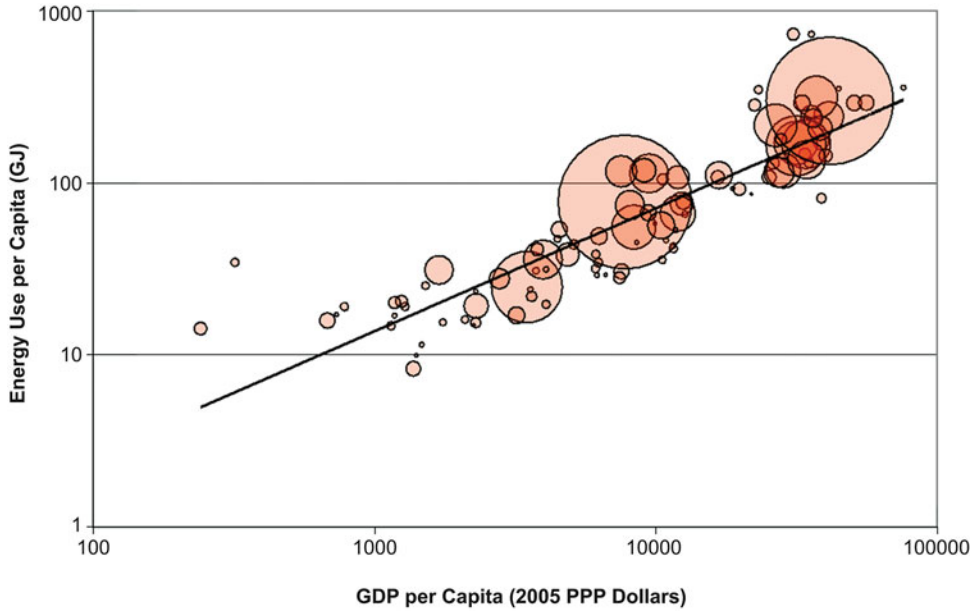
Figure 1 shows that energy use per capita increases with GDP per capita, so that richer countries typically use more energy per person than poorer countries. In fact, this relationship has been very stable over the last several decades, and the graph for 1971 looks very similar except that most countries were poorer and therefore used less energy. Van Benthem (2015) finds similarly that energy intensity – energy used per dollar of GDP – in today's middle-income countries is similar to that in today's developed countries when they were at the same income level.

The slope of the graph is a bit less than 1, so that a 1% increase in income per capita is associated with only a 0.7% increase in energy use per capita (Csereklyei et al. 2016). This means that energy intensity is on average lower in higher income countries, but that this relationship is not so strong (Fig. 2).

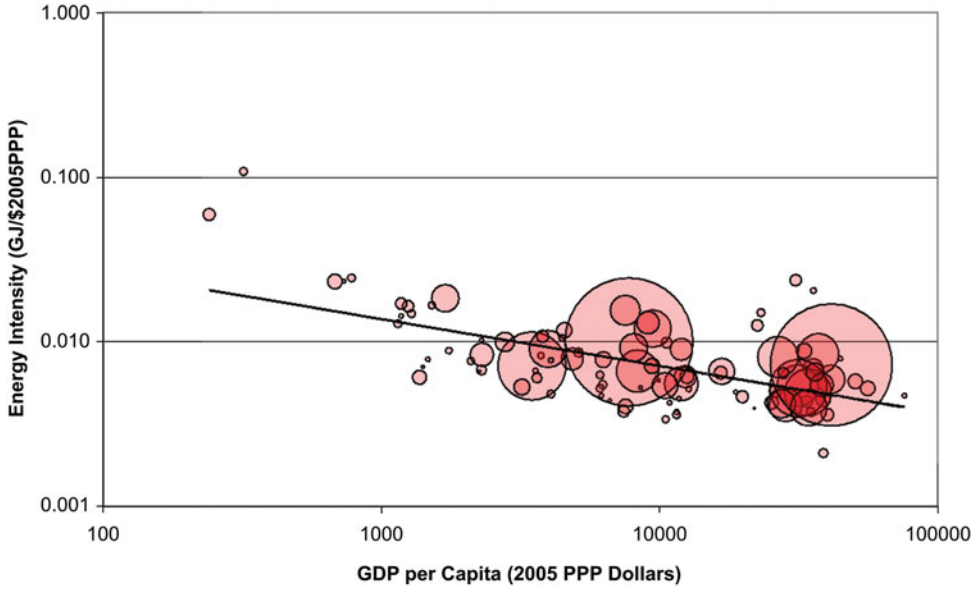
Globally, total energy use has increased over time and energy intensity has decreased (Fig. 3). This is mostly due to countries decreasing in energy intensity as they get richer, as the cross-sectional relationship in Fig. 2 has been fairly stable over time. Energy intensity has, however, become more similar over countries over time, so that countries that were more energy intensive in the 1970s reduced their energy intensity by more than less energy-intensive countries on the whole and the least energy-intensive countries often increased in energy intensity. This convergence relationship is shown in Fig. 4.

Though data are limited to fewer and fewer countries as we go back further in time, these relationships also appear to hold over the last two centuries – energy use has increased, energy intensity has declined globally and countries have converged in energy intensity (Csereklyei et al. 2016).

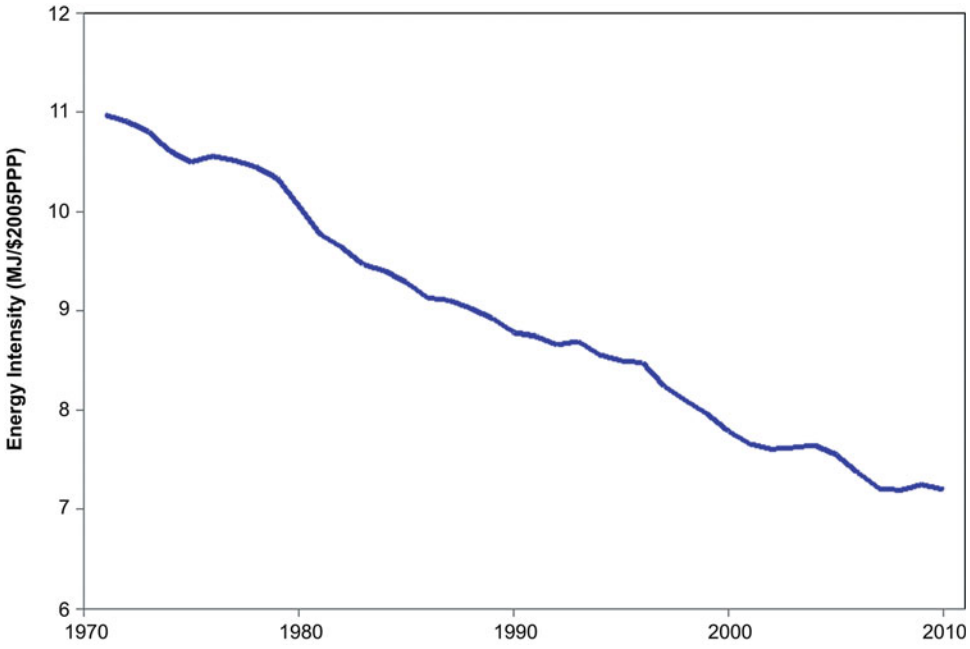
Of course, although energy intensity has declined, per capita energy use has increased over time, and when we also take population growth into account total energy use has risen strongly, though at a slower pace than total world economic output. Between 1971 and 2010 total world energy use increased by about 140% and GDP by 270%. Population increased by 80%.



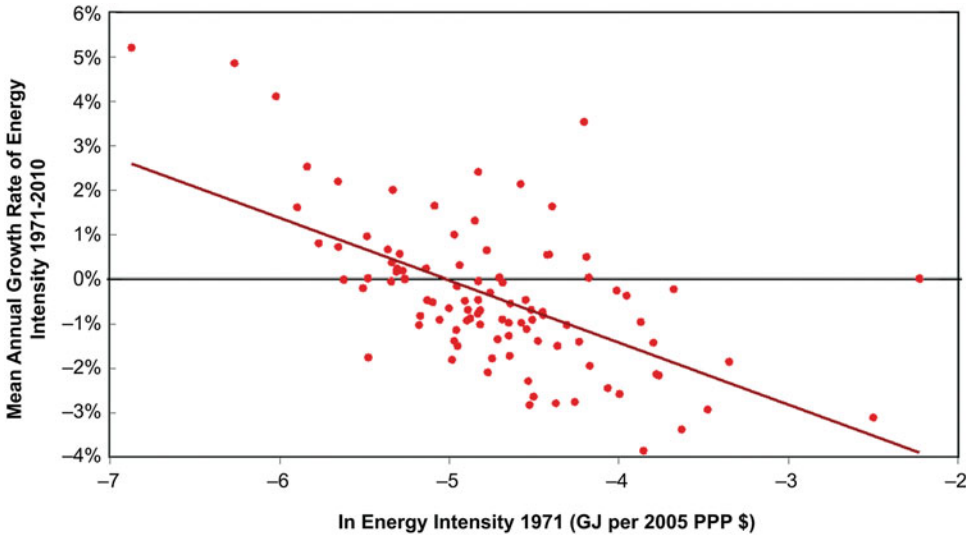
Energy-GDP Relationship, Fig. 1 Energy consumption per capita by GDP per capita 2010. Bubbles are proportional to total energy use. The two largest circles are the USA at upper right and China in the middle (Sources: International Energy Agency and Penn World Table 7.1)



Energy-GDP Relationship, Fig. 2 Energy intensity by GDP per capita 2010. Bubbles are proportional to total energy use. The two largest circles are the USA at lower right and China in the middle (Sources: International Energy Agency and Penn World Table 7.1)



Energy-GDP Relationship, Fig. 3 Global energy intensity (Sources: International Energy Agency and Penn World Table 7.1)



Energy-GDP Relationship, Fig. 4 Convergence of energy intensity (Sources: International Energy Agency and Penn World Table 7.1)

This article next examines the factors that might lead to lower energy intensity with higher GDP and convergence in energy intensity, as seen in Figs. 2, 3, and 4. Then it reviews the literature on the theoretical relationship between energy and economic growth. Third, it reviews estimates of the elasticity of energy use with respect to GDP. The penultimate section looks at the empirical evidence on the question of whether changes in energy use cause changes in

GDP or vice versa. Concluding remarks point to the main gaps in our knowledge.

Factors Affecting the Linkage Between Energy and GDP

Introduction

We saw above that energy intensity is lower in richer countries and has declined globally over time. What are the reasons for this change in the ratio of energy to GDP? We can use a production frontier approach to examine the factors that could weaken or strengthen the linkage between energy use and economic activity over time. A general production frontier, assuming separability between inputs and outputs, is given by:

$$(Q_1,\ldots,Q_m)' = f\left(A_{X1}X_1,\ldots,A_{Xn}X_n,A_{E1}E_1,\ldots,A_{Ep}E_p\right) \tag{1}$$

where the Q_i are various outputs, such as manufactured goods and services, the X_j are various non-energy inputs, such as capital and labour, the E_k are different energy inputs, such as coal and oil, and the A_{Xj} and A_{Ek} are indices of the state of factor-augmenting technology. The relationship between energy and an aggregate of output such as GDP is then affected by:

- substitution between energy and other inputs,
- technological change,
- shifts in the composition of the energy input, and
- shifts in the composition of output.

We discuss each of these below. Also, shifts in the mix of the other inputs – for example to a more capital-intensive economy from a more labour-intensive economy – could affect the relationship between energy and output, but this issue has not been much discussed in the literature and so will not be pursued further here. An important factor offsetting the effects of technological change is the rebound effect, so we discuss this separately. Because all of these factors affect energy intensity, energy intensity is a poor proxy for energy efficiency, which is usually defined more narrowly (Ang 2006; Stern 2012b).

The theoretical rationale for considering energy as a factor of production is discussed later in this article. Of course, if energy is an input to production, then there is also a derived demand function for energy and the level of output or income is one of the factors that determine demand. Estimates of the income elasticity of energy are also discussed below.

Substitutability of Energy and Capital

Koetse et al. (2008) conduct a meta-analysis of the (Morishima) elasticity of substitution (MES) between capital and energy for an increase in the price of energy. Their base case finds that the MES between energy and capital is 0.216, so that capital and energy are poor substitutes. The MES estimated using panel and cross-section data are greater: 0.592 and 0.848, respectively. It is likely that these larger values reflect long-run elasticities and the lower values short-run elasticities (Stern 2012a). Relatively little research has looked at whether capital–energy substitution has driven part of the decline in energy intensity. Stern (2012b) found that capital deepening reduced energy intensity by 7% globally from 1971 to 2007. On the other hand, Wang (2011) found that capital accumulation was the main driver of reduced energy intensity in China.

Energy Efficiency and Technological Change

There are several ways of measuring changes in energy efficiency that take into account the shifts in other factors, such as the quantities of other inputs. The distance function approach measures the change in energy efficiency as the change in the minimum energy requirement to produce a given level of output holding all other inputs constant. Equation (2) compares the minimum energy requirements given the technologies in two different periods but the same levels of inputs and outputs:

$$B_t = \frac{E_0(\mathbf{y}_0,\mathbf{x}_0)}{E_t(\mathbf{y}_0,\mathbf{x}_0)} \tag{2}$$

where $\mathbf{y}$ is the vector of outputs and $\mathbf{x}$ the vector of non-energy inputs with subscripts indicating the

periods and $E_i(\ )$ is a function indicating the minimum energy required in period i in order to achieve the given outputs given the level of inputs. The functions in Equation (2) can be estimated econometrically (e.g. Stern 2012b) or non-parametrically.

A second approach is to use an index of energy augmenting technical change. Based on equation (1), the index of energy augmenting technical change can be constructed as:

$$A_E = \sum_{i=1}^{p} S_i A_{Ei} \tag{3}$$

where S_i are the shares of each type of energy in the total cost of energy. Change over time in this index can be computed using an index method such as Divisia aggregation. The actual energy augmentation indices need to be estimated econometrically.

Bottom up, engineering-based measurements of energy efficiency represent a third approach. For example, Ayres et al. (2003) and Warr et al. (2010) estimate the useful work performed per joule of energy by various fuels and uses of energy. With some exceptions, the general trend over the 20th century in the USA, UK, Japan and Austria has been to greater energy efficiency measured in this way. Over shorter periods energy efficiency has declined in some countries, as it has in the long term for some fuels, especially food and feed.

Estimates of the trends in energy efficiency are mixed (Stern 2011). The direction of change has not been constant and varies across different sectors of the economy. Judson et al. (1999) show that technical innovations tend to introduce more energy-using appliances to households and energy-saving techniques to industry. Stern (2012b) finds that energy efficiency improved from 1971 to 2007 in most developed economies, the former communist countries (including China) and in India. But there was no improvement or a reduction in energy efficiency in many developing economies. Globally, such technological change resulted in a 40% reduction in energy use over the period than would otherwise have been the case and so was the most important driver of reduced energy intensity.

When there is endogenous technological change, changes in prices may induce technological changes. Newell et al. (1999) provide some information on the degree to which energy price increases induce improvements in the energy efficiency of consumer products. For room air conditioners they found that only about one quarter of the gain in energy efficiency since 1973 was induced by higher energy prices. Another quarter was found to be due to raised government standards and labelling. For gas water heaters the induced improvements were close to one half of the total, although much less cost-reducing technical change occurred. Using US data, Popp (2002) similarly finds that increased energy prices have a significant though quantitatively small effect on the rate of patenting in the energy sector. Dechezleprêtre et al. (2011) broaden the analysis to cover patents from 84 national and international patent offices covering various climate mitigation technologies, including renewable energy and energy efficiency technologies. They find that until 1990 patenting in these fields closely followed oil prices. After 1990 patenting increased steadily, although oil prices remained stagnant until 2003. They argue that the increase in patenting since 1990 was driven by environmental policies as they occurred, especially in countries that ratified the Kyoto Treaty.

New energy-using technologies initially diffuse slowly due to high costs of production that are typically lowered radically by a fairly predictable process of learning by doing (Grübler et al. 1999). Diffusion tends to follow a logistic curve, with the speed of diffusion depending, among other things, on how well the innovation fits into the existing infrastructure. Energy-saving innovations such as LED light bulbs would be expected to diffuse rapidly once their price becomes competitive, while more radical innovations that require new support infrastructures diffuse much more slowly due to 'network effects'.

Research also investigates the factors that affect the adoption of energy efficiency policies or energy efficiency technology (Matisoff 2008;

Fredriksson et al. 2004; Gillingham et al. 2009; Linares and Labandeira 2010; Wei et al. 2009; Stern 2012b). Differences in the adoption of energy efficiency technologies across countries and states, over time and among individuals might be optimal due to differences in endowments, preferences or the state of technology. But the rate of adoption may also be inefficient due to market failures and behavioural factors. Market failures include environmental externalities, information problems, liquidity constraints in capital markets, failures of innovation markets and principal–agent problems, such as between landlords and tenants (Gillingham et al. 2009; Linares and Labandeira 2010). Fredriksson et al. (2004) find that the greater the corruptibility of policymakers the less stringent is energy policy, and that the greater lobby group coordination costs are the more stringent energy policy is. Matisoff (2008) finds that the most significant variable affecting the adoption of energy efficiency programs across US states is citizen ideology. A broad band of states from Florida to Idaho had not adopted any policies.

The Rebound Effect

Energy-saving innovations reduce the cost of providing energy services, such as heating, lighting and industrial power. This reduction in cost encourages consumers and firms to use more of the service. As a result, energy consumption usually does not decline by as much as the increase in energy efficiency implies. This difference between the improvement in energy efficiency and the reduction in energy consumption is known as the rebound effect. Rebound effects can be defined for energy-saving innovations in consumption and production. In both cases, the increase in energy use due to increased use of the energy service where an efficiency improvement has happened is called the direct rebound effect. For consumer use of energy, the estimated rebound effects are usually small: in the range of 10–30% (Greening et al. 2000; Sorrell et al. 2009). Roy (2000) argues that because high-quality energy use is still small in households in India, demand is very elastic, and thus rebound effects in the household sector in India and other developing countries can be expected to be larger than in developed economies. Fouquet (2014) confirms how the price elasticity of demand for energy services declines and how the direct rebound effect decreased as Britain developed.

In the case of energy efficiency improvements in industry, the rebound effect at the firm level could be large as the firm could greatly increase its sales as a result of reduced costs. However, under perfect competition for an industry supplying domestic demand it is much harder for the industry as a whole to expand output, so the direct rebound effect would be more limited. Rebound effects are likely to be larger for export industries that have more opportunity to expand production (Grepperud and Rasmussen 2004; Allan et al. 2007; Linares and Labandeira 2010).

As a result of the reduction in the cost of the energy service, consumers will demand less of substitute goods and more of complementary goods. These include other energy services. Firms will make similar changes in their demands for inputs. There will also be additional repercussions throughout the economy. Non-energy goods whose demand has increased require energy in their production. The fall in energy demand may lower the price of energy (Gillingham et al. 2013; Borenstein 2015), increasing energy use again, and the efficiency improvement is a contribution to an increase in total factor productivity, which tends to increase capital accumulation and economic growth, which again results in greater energy usage (Saunders 1992). These additional effects are called indirect rebound effects, though the latter two may be treated separately as 'macro-level rebound effects' (e.g. Howarth 1997). Direct and indirect rebound effects together sum to the economy-wide rebound effect.

Estimates of the economy-wide rebound effect are few in number (e.g. Turner 2009; Barker et al. 2009; Turner and Hanley 2011) and vary widely (Stern 2011; Saunders 2013; Turner 2013). At the economy-wide level, 'backfire', where energy use increases as a result of an efficiency improvement, or even 'super-conservation', where the rebound is negative, are both theoretically possible

(Saunders 2008; Turner 2009). It is usually assumed that the indirect rebound is positive and that the economy-wide rebound will be larger in the long run than in the short run (Saunders 2008). Turner (2013) argues, instead, that because the energy used to produce a dollar's worth of energy is higher than the embodied energy in most other goods, the effect of consumers shifting spending to goods other than energy will mean that the indirect rebound could be negative and the economy-wide rebound may also be negative in the long run. Borenstein (2015) presents further arguments for negative rebounds.

All evidence on the size of the economy-wide rebound effect to date depends on theory-driven models, which have limited empirical validation. Turner (2009) finds that, depending on the assumed values of the parameters in a simulation model, the rebound effect for the UK can range from negative to more than 100%. Barker et al. (2009) provide the only estimate of the global rebound effect, estimating the rebound from a set of IEA recommended energy efficiency policies at 50%.

However, these are rebounds in energy use rather than energy intensity. As the economy-wide rebound effect is largely due to an increase in output, the rebound effect probably has small effects on energy intensity.

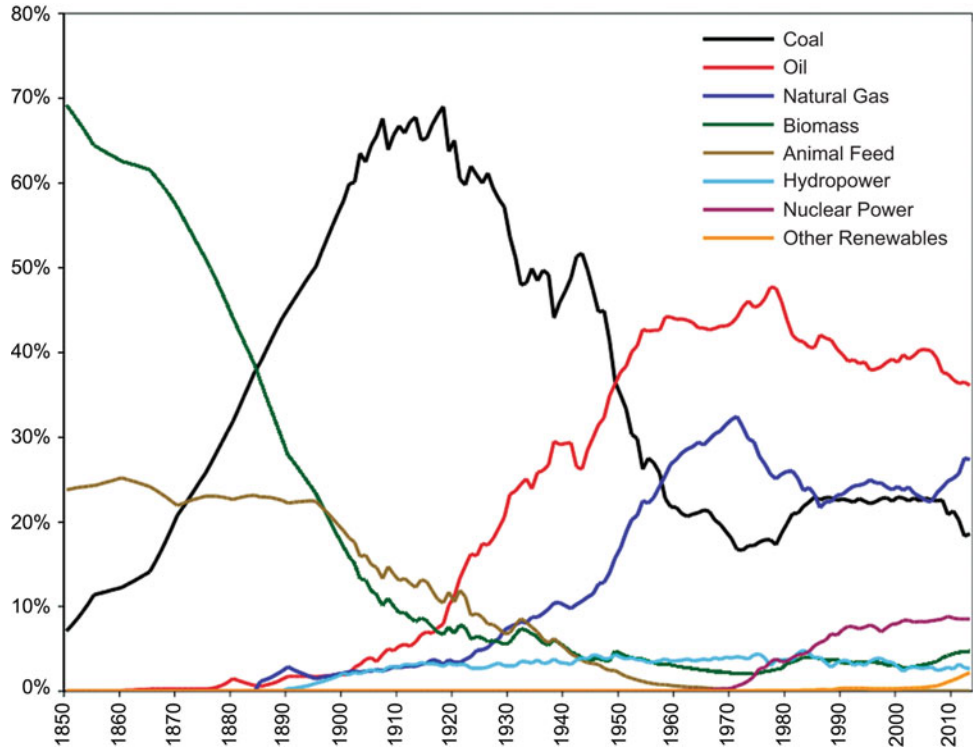
Energy Quality and Shifts in Mix of Energy Inputs

In the course of economic development, countries' fuel mix tends to evolve as the mix of energy sources used shifts to higher quality fuels (Burke 2013). Energy quality is the relative economic usefulness per heat equivalent unit of different fuels and electricity. Fuels have a number of physical attributes that affect their relative qualities, including energy density (heat units per mass unit); power density (rate of heat units produced per unit or per unit time); ease of distribution; the need for a transfer medium; controllability (the ability to direct the position, direction and intensity of energy use); amenability to storage; safety; and environmental impacts (Berndt 1978; Schurr 1982; Cleveland et al. 2000; Stern 2010). Some fuels, in particular electricity, require innovations to allow their use that must be embodied in capital equipment, which can transform the workplace entirely and change work processes, thus contributing to productivity gains (Schurr and Netschert 1960; Toman and Jemelkova 2003; Enflo et al. 2009).

In the least developed economies, as in today's developed economies before the Industrial Revolution, the use of biomass and muscle power dominates. The evolution of the energy mix over the course of economic development and over history in the technologically leading countries depends on each country's endowments of fossil energy and potential for renewables, such as hydroelectricity, but some regularities apply. The share of electricity in total energy use tends to rise. Low-income countries tend to generate electricity from hydropower and oil, while high-income countries have more diverse power sources, including nuclear power. Direct use of coal tends to rise and then fall over time and with income. Natural gas use has increased significantly in recent decades, mostly in more developed economies. Finally, electricity generated from solar and wind power is only now beginning to take off in more developed economies. Figure 5 illustrates this pattern for the USA.

Surprisingly, relatively few studies evaluate the role of the change in energy mix on energy intensity. Schurr and Netschert (1960) were among the first to recognise the economic importance of energy quality in understanding trends in energy and output. Noting that the composition of energy use has changed significantly over time, Schurr and Netschert argued that the general shift to higher quality fuels reduces the amount of energy required to produce a dollar's worth of GDP. Berndt (1990) also noted the key role played by the shifting composition of energy use towards higher quality energy inputs.

Cleveland et al. (1984) and Kaufmann (1992, 2004) presented analyses that explain much of the decline in the US energy/GDP ratio in terms of structural shifts in the economy and shifts from lower to higher quality fuels. Kaufmann (2004) found that shifting away from coal use and, in particular, shifting towards the use of oil reduced



Energy-GDP Relationship, Fig. 5 Composition of US primary energy input 1850–2013 (Source: US Energy Information Administration)

energy intensity in the USA. This shift away from coal more than explained the decline energy intensity over the entire 1929–99 time period. Other studies find, however, a much larger role for technological change than for changes in the composition of energy in the reductions in energy intensity seen around the world. For example, Ma and Stern (2008) find that interfuel substitution had negligible effects on the decline in energy intensity in China between 1994 and 2003. Technological change reduced energy intensity by more than the actual reduction in energy intensity due to the intensity increasing effects of structural change. Stern (2012b) finds that between 1971 and 2007, changes in fuel mix within individual countries increased world energy use by 4%, while global energy intensity declined by 40%. Shifts in the distribution of economic activity towards countries with lower quality energy mixes, such as China and India, contributed further to increasing energy intensity globally.

Shifts in the Composition of Output

Output mix also typically changes over the course of economic development. In the earlier phases of development there is a shift away from agriculture towards heavy industry, while in the later stages of development there is a shift from the more resource-intensive extractive and heavy industrial sectors towards services and lighter manufacturing. Different industries have different energy intensities. It is often argued that this will result in an increase in energy used per unit of output in the early stages of economic development and a reduction in energy used per unit output in the later stages of economic development (Stern 2004).

However, there is reason to believe that the energy-saving effects of structural changes are overstated (Henriques and Kander 2010). When the indirect energy use embodied in manufactured products and services is taken into account, the service and household sectors are more energy-

intensive than they first appear. Service industries still need large energy and resource inputs. The service being sold may be intangible, but the office towers, shopping malls, warehouses, rental apartment complexes etc. where the activity is conducted are very tangible and energy is used in their construction, operation and maintenance. Furthermore, consumers use large amounts of energy and resources in commuting to work, going shopping etc.

Furthermore, on a global scale there may be limits to the extent to which developing countries can replicate the structural shift that has occurred in the developed economies, to the degree that this is due to outsourcing manufacturing overseas rather than simply from an expansion in service activities. However, the evidence shows that trade does not result in reductions in energy use and pollution in developed countries through the offshoring of pollution-intensive industries (Levinson 2010; Aguayo and Gallagher 2005; Kander and Lindmark 2006). Additionally, if the service sector does require substantial material support, it is not clear whether the developed world can continue to shift in the direction of a growing service share of GDP indefinitely. In fact, as manufacturing prices have fallen relative to the prices of services, even the relative decline of manufacturing in developed countries is exaggerated when the relative sizes of the sectors are computed in current prices (Kander 2005).

Kander (2002) and Stern (2012b) find a relatively small role for structural change in reducing energy intensity in Sweden (1800–2000) and the world (1971–2007), respectively. But, using a much finer disaggregation of industries, Sue Wing (2008) finds that structural change explained most of the decline in energy intensity in the USA (1958–2000), especially before 1980.

The Theory of Energy in Economic Production and Growth

Energy as a Factor of Production

Physical laws describe the operating constraints of economic systems (Boulding 1966; Ayres and Kneese 1969). Conservation of mass means that, to obtain a given production output, greater or equal quantities of materials must be used as inputs, and the production process results in residuals or waste (Ayres 1969). Additionally, production requires energy to carry out work to convert materials into desired products and to transport raw materials, goods and people. The second law of thermodynamics (the entropy law) implies that energy cannot be reused and there are limits to how much energy efficiency can be improved. As a result, energy is always an essential factor of production (Stern 1997) and continuous supplies of energy are needed to maintain existing levels of economic activity as well as to grow and develop the economy. Before being used in the production of goods and services, energy and matter must be captured from the environment, and energy must be invested in order to extract useful energy (Hall et al. 1986).

The Mainstream Theory of Growth

Despite these facts, the core mainstream economic growth models disregard energy or other resources. Aghion and Howitt's (2008) textbook on economic growth does discuss growth and the environment, but only in a chapter near the end of the book. Acemoglu's (2009) textbook does not cover the topic at all. There has been some analysis of the potential for resources to constrain growth in the journal literature, but it has mostly been contained within the sub-field of environmental and resource economics, and the main focus has been on the implications of non-renewable resources for economic growth.

Solow (1974) introduced non-renewable resources – which could represent fossil or nuclear fuels – into neoclassical growth models and showed that sustainability – or the ability of a nation to support a constant level of economic production indefinitely – is achievable under certain institutional and technical conditions. Assuming that there is no population growth or technological progress, Solow shows that technology must allow the use of natural resources and manufactured capital – machines and buildings – to be sufficiently responsive to changes in prices. As the price of natural resources relative to that of capital rises, capital is substituted for resources in

production. In Solow's (1974) model the elasticity of substitution is 1, as implied by the Cobb–Douglas production function. This means that resources are essential, but that a constant level of production could be maintained even with infinitesimally small resource inputs. An elasticity of substitution greater than unity means that resources are not essential and so achieving sustainability is much easier. These are all conditions concerning the technology available to society. But the institutional framework – for example, whether an economy is a free market economy or whether it follows a particular planning rule – is just as important. From an institutional perspective, sustainability can be achieved only if the welfare of future generations is given equal weight to that of the present generation. This implies that the discount rate used to aggregate costs and benefits over time must be zero.

If instead the economy is a free market economy with perfect competition, but has the same technology as Solow's (1974) model, the resources are exhausted and consumption and social welfare eventually fall to zero (Stiglitz 1974a). Dasgupta and Heal (1979) showed that with any constant discount rate the efficient growth path also leads to eventual depletion of the natural resource endowment and the collapse of the economy. Hartwick (1977, 1995) has shown that, if sustainability is technologically feasible, a constant level of consumption can be achieved by investing the rents from exhaustible resources in other forms of capital, which in turn can substitute for exhausted resources. It is difficult to apply this rule in practice, as the rents and capital must be valued at prices that are compatible with sustainability (Asheim 1994; Asheim et al. 2003; Pezzey 2004). Such prices are unknowable given that we have poor understanding of the costs of current environmental damage and resource depletion or of the future development of technology.

In addition to the substitution of capital for resources, technological change might permit continued growth or at least constant consumption in the face of a finite resource base. Stiglitz (1974b) showed that, when the elasticity of substitution between capital and resources is 1, exogenous technical progress will allow consumption to grow over time if the rate of technological change divided by the discount rate is greater than the output elasticity of resources. Technological change might enable sustainability, even with an elasticity of substitution of less than 1.

Once again, technical feasibility does not guarantee sustainability. Depending on preferences for current versus future consumption, technological change might instead result in faster depletion of the resource (Smulders 2005).

The Ecological Economics Approach

A prominent tradition in ecological economics, known as the biophysical economics approach (Hall et al. 1986), is based on thermodynamics (Georgescu-Roegen 1971; Costanza 1980; Cleveland et al. 1984; Hall et al. 1986, 2003; Ayres and Warr 2005, 2009; Murphy and Hall 2010). Ecological economists usually argue that substitution between capital and resources can only play a limited role in mitigating the scarcity of resources (Stern 1997). Furthermore, some ecological economists downplay the role of technological change in productivity growth, arguing that growth is a result of either increased energy use or innovations allowing the more productive use of energy (Hall et al. 1986, 2003; Cleveland et al. 1984; Ayres and Warr 2009). Therefore, in this view, increased energy use is the main or only cause of economic growth.

In this approach, value is derived from the action of energy that is directed by capital and labour. Energy flows into the economy from fossil fuels and the Sun.

In some biophysical economic models, geological constraints fix the rate of energy extraction so that the flow rather than the stock can be considered as the primary input to production (Gever et al. 1986). Capital and labour are considered as intermediate inputs that are created and maintained by the primary input of energy and flows of matter. The level of the flows is computed in terms of the embodied energy use associated with them. Prices of goods should then ideally be determined by their embodied energy cost (Hannon 1973) – a normative energy theory of

value – or are seen as actually being correlated with energy cost (Costanza 1980) – a positive energy theory of value (Common 1995). This theory – like the Marxian paradigm – must then explain how labour, capital etc. end up receiving part of the surplus (Kaufmann 1987; Burkett 2003; Hornborg 2014).

However, because the quality of resources and the level of technology do affect the amount of energy needed to produce goods and services, it is difficult to argue for a model where energy is the sole factor of production (Stern 1999). For example, the quality of resources such as oil reservoirs is critical in determining the energy required to extract and process fuels. As an oil reservoir is depleted, the energy needed to extract oil increases. On the positive side, improved geophysical knowledge and techniques can increase the extent to which oil can be extracted for a given energy cost. Odum's energy approach (Brown and Herendeen 1996) and the framework developed by Costanza (1980) address the resource quality issue by including the solar and geological energy embodied in natural resource inputs in indicators of total embodied energy. An alternative approach is to measure material and energy inputs on the common basis of their exergy (Ayres et al. 1998; Ukidwe and Bakshi 2007).

However, both approaches seem too reductionist. For example, other services provided by nature, such as nutrient recycling, the provision of clean air and water, pollination and the climate system, that make economic production – and life itself – possible should also then be accounted for. Models that allow a number of different factors of production while complying with the physical laws of the conservation of mass and thermodynamics to varying degrees were developed by Georgescu-Roegen (1971), Perrings (1987), and O'Connor (1993) among others. The ecological economics approach does not have to reduce to an energy-only model of the economy.

A key concept in biophysical economics is energy return on investment (EROI), which is the ratio of useful energy produced by an energy supply system to the amount of energy invested in extracting that energy. Lower quality energy resources have lower EROIs. Biophysical economists argue that the more energy that is required to extract energy, the less energy is available for other uses and the poorer an economy will be. In this view, the increase in EROI allowed by the switch from biomass to fossil fuels enabled the Industrial Revolution and the period of modern economic growth that followed it (Hall et al. 1986).

Thus, declining EROI would threaten not just growth but overall economic output and, therefore, sustainability. Murphy and Hall (2010) document EROI for many energy sources, arguing that it is declining over time despite the extensive innovation in the industry. Wind and direct solar energy have more favourable EROIs than biomass fuels, but worse than most fossil fuels. However, unlike fossil fuels, the EROI of these energy sources tends to improve over time due to innovation (Kubiszewski et al. 2010). Declining EROI could be mitigated by substituting other inputs for energy or by improving the efficiency with which energy is used. However, biophysical economists argue that both these processes have limits.

Substitution can occur *within* a category of similar production inputs – for example between different fuels – and *between* different categories of inputs – for example between energy and machines. There is also a distinction to be made between substitution at the micro level – for example within a single engineering process or at a single firm – and at the macro level – in the economy as a whole.

As shown in Fig. 5 for the USA, the long-run pattern of energy use in industrial economies has been dominated by substitutions from wood and animal power to coal, oil, natural gas and primary electricity (Grübler et al. 2012). Meta-analysis of existing studies of interfuel substitution suggests that the long-run substitution possibilities at the level of the industrial sector as a whole are good. But there seems to be less substitutability at the macro-economic level (Stern 2012a).

Ecological economists emphasise the importance of limits to inter-category substitution – in particular, the substitution of manufactured capital for resources including energy (Costanza and Daly 1992). Thermodynamic limits on substitution can be approximated by a production function with an elasticity of substitution significantly below one (Stern 1997). As discussed above, a

meta-analysis of the existing empirical literature finds that the elasticity of substitution between capital and energy is less than 1 but much greater than 0 (Koetse et al. 2008).

In addition to this micro-economic limit to substitution, there may also be macroeconomic limits to substitution. The construction, operation and maintenance of tools, machines and factories require a flow of materials and energy. Similarly, the humans that direct manufactured capital consume energy and materials. Thus, producing more of the 'substitute' for energy – manufactured capital – requires more of the thing that it is supposed to substitute for. This again limits potential substitutability (Cleveland et al. 1984).

The mainstream economic argument that technological change can overcome limited substitutability would be more convincing if technological change were really something different from substitution. Changes in technology occur when new techniques are developed. However, these new techniques represent the substitution of knowledge for other factors of production. The knowledge is embodied in improved capital goods and more skilled workers and managers. But there are still thermodynamic restrictions on the extent to which energy and material flows can be reduced in this way. Although knowledge is non-rival in use, it must be used in conjunction with the other inputs, such as energy, and the productivity of knowledge is limited by the available quantities of those inputs.

Synthesis: Unified Model of Energy and Growth

As a first step to integrating the ecological economic and mainstream approaches and explaining historical economic growth, Stern and Kander (2012) add an energy input that has low substitutability with capital and labour to Solow's (1956) growth model. As discussed above, low substitutability between capital and energy is one of the key assumptions of ecological economists. Using 200 years of Swedish data Stern and Kander estimate that the elasticity of substitution between energy and the other two inputs is 0.65. This figure is similar to the other estimates of the elasticity also discussed in the previous section. Stern and Kander ignore the issue of whether the energy resource is non-renewable, as depletion of fossil fuels does not seem to have been a very important factor in constraining economic growth to date.

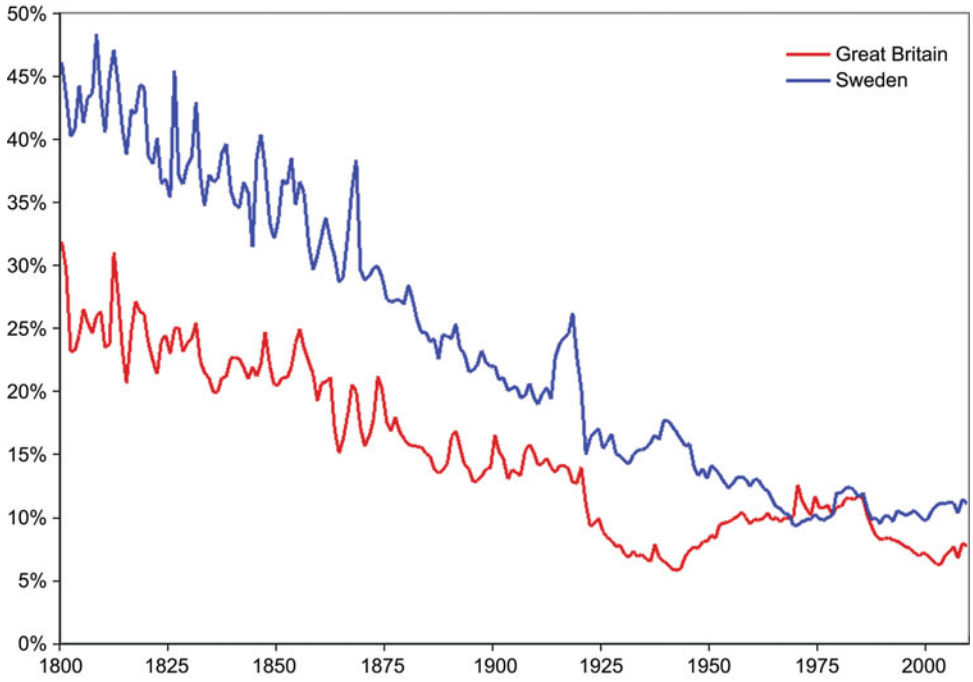
Assuming that the elasticity of substitution between energy and capital is less than 1 allows the share of energy in production costs to fall over time. When the elasticity of substitution is unity, cost shares must be constant in the long run. The cost share of energy has fallen in the long run in both Britain and Sweden, countries for which we have data from 1800 till the present (Fig. 6). An elasticity of substitution of less than unity also allows us to distinguish between labour-augmenting innovations and energy-augmenting innovations, which again is not possible using a Cobb–Douglas production function.

The production function is given by:

$$Y = \left[(1-\gamma)\left(A_L^{\beta}L^{\beta}K^{1-\beta}\right)^{\phi} + \gamma(A_E E)^{\phi}\right]^{\frac{1}{\phi}} \quad (4)$$

Equation (4) embeds a Cobb–Douglas function of capital, K, and labour, L, $L^{\beta}K^{1-\beta}$, in a constant elasticity of substitution production function of this combined input and energy, E, to produce gross output, Y. $\phi = (\sigma - 1)/\sigma$, where σ is the elasticity of substitution between energy and the capital–labour aggregate. A_L and A_E are the augmentation indices of labour and energy, which can be interpreted as reflecting both changes in technology that augment the effective supply of the factor in question and changes in the quality of the respective factors. $A_E E$ and $A_L L$ are called effective energy and effective labour, respectively.

In Solow's (1956) model, as long as there is technological change the economy can grow. In Stern and Kander's model, depending on the availability of energy and the nature of technological change, energy can be either a constraint on growth or an enabler of growth. When effective energy, $A_E E$, is very abundant, the model behaves very similarly to Solow's original model and energy neither constrains nor drives growth. The more energy there is, the less important energy appears to be. But when effective energy is relatively scarce,



Energy-GDP Relationship, Fig. 6 Share of energy in total production costs: Britain and Sweden 1800–2009 (Source: Gentvilaite et al. (2015))

the level of output depends on the level of energy supply and the level of energy-augmenting technology. Labour-augmenting technological change alone no longer results in economic growth.

Before the Industrial Revolution most energy was in the form of wood and animal and human muscle power – wind- and water-power contributed relatively little energy (Kander et al. 2014). The supply of this renewable energy was constrained by the availability of land, so energy was scarce (Wrigley 2010). Therefore, as the data show (Maddison 2001), until the Industrial Revolution, output per capita was generally low and economic growth was not sustained. Stern and Kander (2012) find that increases in energy use and energy-augmenting technological change were the main contributors to economic growth in the 19th and early 20th centuries, but in the second half of the 20th century labour-augmenting technological change became the main driver of growth in income per capita, as it is in the Solow growth model.

The Elasticity of Energy with Respect to GDP

How much does energy use increase with economic growth? Various studies have estimated by how much energy use tends to be higher as income increases without controlling for other factors, while other studies attempt to estimate the macro-level elasticity of demand for energy, controlling for energy prices and changes in technology. I summarise some recent econometric results. As mentioned in the introduction, Csereklyei et al. (2016) find that there has been a remarkably stable relationship between energy and GDP over the last four decades in a sample of 99 countries. The elasticity, which does not control for energy prices or technological change, is 0.7. Similarly, using panel data for middle-income countries, including today's developed countries in earlier decades, Van Benthem (2015) finds an elasticity of 0.9 or 0.97 controlling for energy prices and time effects. He finds lower elasticities for higher and lower income

bands. Similarly, Fouquet (2014) finds that energy income elasticities first rose and then fell over the course of economic development in Britain. Using a cointegration approach, Joyeux and Ripple (2011) estimate that the long-run income elasticity is 1.08 for OECD countries and 0.853 for 19 developing countries between 1973 and 2007. These estimates do not control for energy prices or time effects. Thus energy is a normal good and most estimates of the elasticity are between 0.5 and unity. However, there does not seem to be a consensus on whether the income elasticity declines or not with increasing income.

Testing for Causality Between Energy and GDP

Two methods for testing for causality among time series variables are Granger causality tests and cointegration analysis (Granger 1969; Engle and Granger 1987). Hendry and Juselius (2000) discuss the application of these methods to energy economics, where they have been applied extensively to test for causality and cointegration between energy, GDP and other variables from the late 1970s on (Kraft and Kraft 1978; Ozturk 2010). There are now hundreds of journal articles on this topic (Bruns et al. 2014).

Early studies relied on Granger causality tests on unrestricted vector autoregressions (VARs) in levels of the variables, while more recent studies use cointegration methods. A vector autoregression model consists of one regression equation for each variable of interest in a system. Each variable is regressed on lagged values of itself and all other variables in the system. If the coefficients of the lagged values of variable X in the equation for dependent variable Y are jointly statistically significant, then X is said to Granger cause Y. Cointegration analysis tests whether variables that have stochastic trends – their trend is a random walk – share a common trend. If so, then at least one variable must Granger cause the other.

Early studies also used bivariate models of energy and output, while more recent research tends to employ multivariate models. Ignoring other relevant variables can generate spurious causality findings. The most common additional variables used are capital and labour or energy prices. A third way to differentiate among models is whether energy is measured in standard heat units or whether a method is used to account for differences in quality among fuels.

The results of early studies that tested for Granger causality using a bivariate model were generally inconclusive (Stern 1993). Stern tested for Granger causality in the USA in a multivariate setting using a vector autoregression (VAR) model of GDP, capital and labour inputs, and a Divisia index of quality-adjusted energy use in place of the usual heat equivalent of energy use. When both the multivariate approach and quality-adjusted energy index were employed, energy use was found to Granger cause GDP.

Yu and Jin (1992) conducted the first cointegration study of the energy–GDP relationship using the bivariate approach. Stern (2000) estimated a dynamic cointegration model for GDP, quality weighted energy, labour and capital. The analysis showed that there is a cointegrating relation between the four variables and, depending on the version of the model used, found that energy Granger causes GDP or that there is mutual causation between energy and GDP. Some subsequent research appeared to confirm these findings using other measures of energy quality (Warr and Ayres 2010) or data for other countries (Oh and Lee 2004; Ghali and El-Sakka 2004) and panels of many countries (Lee and Chang 2008; Lee et al. 2008).

Bruns et al. (2014) carry out a meta-analysis of 75 single-country Granger causality and cointegration studies comprising more than 500 tests of causality in each direction. They find that most seemingly statistically significant results in the literature are probably the result of statistical biases that occur in models that use short time series of data – 'overfitting bias' – or the result of the selection for publication of statistically significant results – 'publication bias'. The most robust findings in the literature are that growth causes energy use when energy prices are controlled for in the underlying studies. Using a panel cointegration model of GDP, energy use and energy prices for 26 OECD countries (1978–2005), Costantini and Martini (2010) also

find that in the long run GDP growth drives energy use and energy prices, though in the short run energy prices cause GDP and energy use, and energy use and GDP are mutually causative.

However, Bruns et al. (2014) find that studies that control for capital do not find a genuine effect of energy on growth or vice versa. But they had too small a number of studies that used quality-adjusted energy to test whether there was a genuine relationship between energy and growth when this measure of energy use was employed. So their findings do not necessarily contradict the previous research by Stern and others reviewed above.

Gaps in Knowledge

As this article has shown, the relationship between energy and GDP is one where there is remarkably little consensus, and large gaps in knowledge remain. The field of energy economics has expanded rapidly in the last decade, but much research is repetitive and adds little to existing knowledge (Smyth and Narayan 2015). In particular, there is a very large literature using reduced form time series models to test for causality and cointegration between energy and output. But this literature is completely inconclusive, with equal numbers of studies finding causation in each direction (Bruns et al. 2014). Research in this area needs to be more closely based on testing potential mechanisms which link energy and output. However, researchers are only starting to build theoretical models of the role of energy in the economic growth process.

There is also a lack of consensus in research on the drivers of changes in energy intensity. In particular, energy intensity has risen in many developing countries. The reasons for this are little researched. There is also a lack of consensus on the size of the economy-wide rebound effect. Existing estimates are all derived from simulation models and range from negative rebound to backfire, where energy efficiency improvements actually increase rather than reduce energy use. Therefore there is little guidance on the potential for energy efficiency policies to actually conserve energy.

Research is also hampered by inadequate data. With the exception of traditional biomass, energy use data are normally of good quality. But data on prices is much more fragmentary. Most economic research is based on understanding the linkages between prices and quantities. So this is an important area where international comparable datasets could be very useful.

See Also

- ▶ Causality in Economics and Econometrics
- ▶ CES Production Function
- ▶ Ecological Economics
- ▶ Economic Growth
- ▶ Energy Economics
- ▶ Granger–Sims Causality
- ▶ Rebound Effects

Bibliography

Acemoglu, D. 2009. *Introduction to economic growth*. Princeton: Princeton University Press.

Aghion, P., and P. Howitt. 2008. *The economics of growth*. Cambridge, MA: MIT Press.

Aguayo, F., and K.P. Gallagher. 2005. Economic reform, energy, and development: The case of Mexican manufacturing. *Energy Policy* 33: 829–837.

Allan, G., N. Hanley, P. McGregor, K. Swales, and K. Turner. 2007. The impact of increased efficiency in the industrial use of energy: A computable general equilibrium analysis for the United Kingdom. *Energy Economics* 29: 779–798.

Ang, B.W. 2006. Monitoring changes in economy-wide energy efficiency: From energy- GDP ratio to composite efficiency index. *Energy Policy* 34: 574–582.

Asheim, G.B. 1994. Net national product as an indicator of sustainability. *Scandinavian Journal of Economics* 96: 257–265.

Asheim, G.B., W. Buchholz, and C. Withagen. 2003. The Hartwick rule: Myths and facts. *Environmental and Resource Economics* 25: 129–150.

Ayres, R.U., and A.V. Kneese. 1969. Production, consumption and externalities. *American Economic Review* 59: 282–297.

Ayres, R.U., and B. Warr. 2005. Accounting for growth: The role of physical work. *Structural Change and Economic Dynamics* 16: 181–209.

Ayres, R.U., and B. Warr. 2009. *The economic growth engine: How energy and work drive material prosperity*. Cheltenham: Edward Elgar.

Ayres, R.U., L.W. Ayres, and K. Martinás. 1998. Exergy, waste accounting, and life-cycle analysis. *Energy* 23 (5): 355–363.

Ayres, R.U., L.W. Ayres, and B. Warr. 2003. Exergy, power and work in the US economy, 1900–1998. *Energy* 28: 219–273.

Barker, T., A. Dagoumas, and J. Rubin. 2009. The macroeconomic rebound effect and the world economy. *Energy Efficiency* 2: 411–427.

Berndt, E.R. 1978. Aggregate energy, efficiency, and productivity measurement. *Annual Review of Energy* 3: 225–273.

Berndt, E.R. 1990. Energy use, technical progress and productivity growth: A survey of economic issues. *The Journal of Productivity Analysis* 2: 67–83.

Borenstein, S. 2015. A microeconomic framework for evaluating energy efficiency rebound and some implications. *Energy Journal* 36 (1): 1–21.

Boulding, K. 1966. The economics of the coming spaceship Earth. In *Environmental quality in a growing economy*, ed. H. Jarett. Baltimore: Johns Hopkins University Press.

Brown, M.T., and R.A. Herendeen. 1996. Embodied energy analysis and emergy analysis: A comparative view. *Ecological Economics* 19: 219–236.

Bruns, S.B., C. Gross, and D.I. Stern. 2014. Is there really Granger causality between energy use and output? *Energy Journal* 35 (4): 101–134.

Burke, P.J. 2013. The national-level energy ladder and its carbon implications. *Environment and Development Economics* 18 (4): 484–503.

Burkett, P. 2003. The value problem in ecological economics: Lessons from the physiocrats and Marx. *Organization & Environment* 16 (2): 137–167.

Cleveland, C.J., R. Costanza, C.A.S. Hall, and R.K. Kaufmann. 1984. *Energy and the U.S. economy: A biophysical perspective*. *Science* 225: 890–897.

Cleveland, C.J., R.K. Kaufmann, and D.I. Stern. 2000. Aggregation and the role of energy in the economy. *Ecological Economics* 32: 301–318.

Common, M.S. 1995. *Sustainability and policy: Limits to economics*. Melbourne: Cambridge University Press.

Costantini, V., and C. Martini. 2010. The causality between energy consumption and economic growth: A multisectoral analysis using non-stationary cointegrated panel data. *Energy Economics* 32: 591–603.

Costanza, R. 1980. Embodied energy and economic valuation. *Science* 210: 1219–1224.

Costanza, R., and H.E. Daly. 1992. Natural capital and sustainable development. *Conservation Biology* 6: 37–46.

Csereklyei, Z., M.d.M. Rubio Varas, and D.I. Stern. 2016. Energy and economic growth: The stylized facts. *Energy Journal* 37 (2): 223–255.

Dasgupta, P.S., and G.M. Heal. 1979. *Economic theory and exhaustible resources*. Cambridge: Cambridge University Press.

Dechezleprêtre, A., M. Glachant, I. Haščič, N. Johnstone, and Y. Ménière. 2011. Invention and transfer of climate change-mitigation technologies: A global analysis. *Review of Environmental Economics and Policy* 5 (1): 109–130.

Enflo, K., A. Kander, and L. Schön. 2009. *Electrification and energy productivity. Ecological Economics 68*: 2808–2817.

Engle, R.E., and C.W.J. Granger. 1987. Cointegration and error-correction: Representation, estimation, and testing. *Econometrica* 55: 251–276.

Fouquet, R. 2014. Long run demand for energy services: Income and price elasticities over 200 years. *Review of Environmental Economics and Policy* 8 (2): 186–207.

Fredriksson, P.G., H.R.J. Vollebergh, and E. Dijkgraaf. 2004. Corruption and energy efficiency in OECD countries: Theory and evidence. *Journal of Environmental Economics and Management* 47: 207–231.

Gentvilaite, R., A. Kander, and P. Warde. 2015. The role of energy quality in shaping long-term energy intensity in Europe. *Energies* 8 (1): 133–153.

Georgescu-Roegen, N. 1971. *The entropy law and the economic process*. Cambridge, MA: Harvard University Press.

Gever, J., R.K. Kaufmann, D. Skole, and C. Vorosmarty. 1986. *Beyond oil: The threat to food and fuel in the coming decades*. Cambridge, MA: Ballinger.

Ghali, K.H., and M.I.T. El-Sakka. 2004. Energy use and output growth in Canada: A multivariate cointegration analysis. *Energy Economics* 26: 225–238.

Gillingham, K., R.G. Newell, and K. Palmer. 2009. Energy efficiency economics and policy. *Annual Review of Resource Economics* 1: 597–620.

Gillingham, K., M.J. Kotchen, D.S. Rapson, and G. Wagner. 2013. The rebound effect is overplayed. *Nature* 493: 475–476.

Granger, C.W.J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37: 424–438.

Greening, L.A., D.L. Greene, and C. Difiglio. 2000. Energy efficiency and consumption – The rebound effect – A survey. *Energy Policy* 28: 389–401.

Grepperud, S., and I. Rasmussen. 2004. A general equilibrium assessment of rebound effects. *Energy Economics* 26: 261–282.

Grübler, A., N. Nakicenovic, and D.G. Victor. 1999. Dynamics of energy technologies and global change. *Energy Policy* 27: 247–280.

Grübler, A., T.B. Johansson, L. Mundaca, N. Nakicenovic, S. Pachauri, K. Riahi, H.-H. Rogner, and L. Strupeit. 2012. Chapter 1. Energy primer. In *Global energy assessment – toward a sustainable future*. Cambridge, UK/New York/Laxenburg: Cambridge University Press/International Institute for Applied Systems Analysis.

Hall, C.A.S., C.J. Cleveland, and R.K. Kaufmann. 1986. *Energy and resource quality: The ecology of the economic process*. New York: Wiley Interscience.

Hall, C.A.S., P. Tharakan, J. Hallock, C.J. Cleveland, and M. Jefferson. 2003. Hydrocarbons and the evolution of human culture. *Nature* 426: 318–322.

Hannon, B. 1973. An energy standard of value. *Annals of the American Academy* 410: 139–153.

Hartwick, J.M. 1977. Intergenerational equity and the investing of rents from exhaustible resources. *American Economic Review* 66: 972–974.

Hartwick, J.M. 1995. Constant consumption paths in open economies with exhaustible resources. *Review of International Economics* 3: 275–283.

Hendry, D.F., and K. Juselius. 2000. Explaining cointegration analysis: Part 1. *Energy Journal* 21 (1): 1–42.

Henriques, S.T., and A. Kander. 2010. The modest environmental relief resulting from the transition to a service economy. *Ecological Economics* 70: 271–282.

Hornborg, A. 2014. Ecological economics, Marxism, and technological progress: Some explorations of the conceptual foundations of theories of ecologically unequal exchange. *Ecological Economics* 105: 11–18.

Howarth, R.B. 1997. Energy efficiency and economic growth. *Contemporary Economic Policy* 25: 1–9.

Joyeux, R., and R.D. Ripple. 2011. Energy consumption and real income: A panel cointegration multi-country study. *Energy Journal* 32 (2): 107–141.

Judson, R.A., R. Schmalensee, and T.M. Stoker. 1999. Economic development and the structure of demand for commercial energy. *Energy Journal* 20 (2): 29–57.

Kander, A. 2002. *Economic growth, energy consumption and CO_2 emissions in Sweden 1800–2000.* Lund studies in economic history, vol. 19. Lund.

Kander, A. 2005. Baumol's disease and dematerialization of the economy. *Ecological Economics* 55 (1): 119–130.

Kander, A., and M. Lindmark. 2006. Foreign trade and declining pollution in Sweden: A decomposition analysis of long-term structural and technological effects. *Energy Policy* 34 (13): 1590–1599.

Kander, A., P. Malanima, and P. Warde. 2014. *Power to the people – Energy and economic transformation of Europe over four centuries*. Princeton: Princeton University Press.

Kaufmann, R.K. 1987. Biophysical and Marxist economics: Learning from each other. *Ecological Modelling* 38: 91–105.

Kaufmann, R.K. 1992. A biophysical analysis of the energy/real GDP ratio: Implications for substitution and technical change. *Ecological Economics* 6: 35–56.

Kaufmann, R.K. 2004. The mechanisms for autonomous energy efficiency increases: A cointegration analysis of the US energy/GDP ratio. *Energy Journal* 25 (1): 63–86.

Koetse, M.J., H.L.F. de Groot, and R.J.G.M. Florax. 2008. Capital–energy substitution and shifts in factor demand: A meta-analysis. *Energy Economics* 30: 2236–2251.

Kraft, J., and A. Kraft. 1978. On the relationship between energy and GNP. *Journal of Energy and Development* 3: 401–403.

Kubiszewski, I., C.J. Cleveland, and P.K. Endres. 2010. Meta-analysis of net energy return for wind power systems. *Renewable Energy* 35: 218–225.

Lee, C.-C., and C.-P. Chang. 2008. Energy consumption and economic growth in Asian economies: A more comprehensive analysis using panel data. *Resource and Energy Economics* 30 (1): 50–65.

Lee, C.-C., C.-P. Chang, and P.-F. Chen. 2008. Energy-income causality in OECD countries revisited: The key role of capital stock. *Energy Economics* 30: 2359–2373.

Levinson, A. 2010. Offshoring pollution: Is the United States increasingly importing polluting goods? *Review of Environmental Economics and Policy* 4 (1): 63–83.

Linares, P., and X. Labandeira. 2010. Energy efficiency: Economics and policy. *Journal of Economic Surveys* 24 (3): 583–592.

Ma, C., and D.I. Stern. 2008. China's changing energy intensity trend: A decomposition analysis. *Energy Economics* 30 (3): 1037–1053.

Maddison, A. 2001. *The world economy: A millennial perspective*. *Paris*: *OECD*.

Matisoff, D.C. 2008. The adoption of state climate change policies and renewable portfolio standards: Regional diffusion or internal determinants? *Review of Policy Research* 25 (6): 527–546.

Murphy, D.J., and C.A.S. Hall. 2010. Year in review – EROI or energy return on (energy) invested. *Annals of the New York Academy of Sciences* 1185: 102–118.

Newell, R.G., A.B. Jaffe, and R.N. Stavins. 1999. The induced innovation hypothesis and energy-saving technological change. *Quarterly Journal of Economics* 114: 941–975.

O'Connor, M.P. 1993. Entropic irreversibility and uncontrolled technological change in the economy and environment. *Journal of Evolutionary Economics* 34: 285–315.

Oh, W., and K. Lee. 2004. Causal relationship between energy consumption and GDP revisited: The case of Korea 1970–1999. *Energy Economics* 26: 51–59.

Ozturk, I. 2010. A literature survey on energy–growth nexus. *Energy Policy* 38: 340–349.

Perrings, C.A. 1987. *Economy and environment: A theoretical essay on the interdependence of economic and environmental systems*. *Cambridge*: *Cambridge University Press*.

Pezzey, J.C.V. 2004. Sustainability tests with amenities, and change in technology, trade and population. *Journal of Environmental Economics and Management* 48: 613–631.

Popp, D. 2002. Induced innovation and energy prices. *American Economic Review* 92: 160–180.

Roy, J. 2000. The rebound effect: Some empirical evidence from India. *Energy Policy* 28: 433–438.

Saunders, H.D. 1992. The Khazzoom–Brookes postulate and neoclassical growth. *Energy Journal* 13 (4): 131–148.

Saunders, H.D. 2008. Fuel conserving (and using) production functions. *Energy Economics* 30: 2184–2235.

Saunders, H.D. 2013. Historical evidence for energy efficiency rebound in 30 US sectors and a toolkit for rebound analysts. *Technological Forecasting and Social Change* 80: 1317–1330.

Schurr, S. 1982. Energy efficiency and productive efficiency: Some thoughts based on American experience. *Energy Journal* 3 (3): 3–14.

Schurr, S., and B. Netschert. 1960. *Energy and the American economy, 1850–1975*. Baltimore: Johns Hopkins University Press.

Smulders, S. 2005. Endogenous technical change, natural resources and growth. In *Scarcity and growth in the new millennium*, ed. R. Ayres, D. Simpson, and M. Toman. Washington, DC: Resources for the Future.

Smyth, R., and P.K. Narayan. 2015. Applied econometrics and implications for energy economics research. *Energy Economics* 50: 351–358.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.

Solow, R.M. 1974. Intergenerational equity and exhaustible resources. *Review of Economic Studies* 41 (5): 29–46.

Sorrell, S., J. Dimitropoulos, and M. Sommerville. 2009. Empirical estimates of the direct rebound effect: A review. *Energy Policy* 37: 1356–1371.

Stern, D.I. 1993. Energy use and economic growth in the USA: A multivariate approach. *Energy Economics* 15: 137–150.

Stern, D.I. 1997. Limits to substitution and irreversibility in production and consumption: A neoclassical interpretation of ecological economics. *Ecological Economics* 21: 197–215.

Stern, D.I. 1999. Is energy cost an accurate indicator of natural resource quality? *Ecological Economics* 31: 381–394.

Stern, D.I. 2000. A multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Economics* 22: 267–283.

Stern, D.I. 2004. The rise and fall of the environmental Kuznets curve. *World Development* 32 (8): 1419–1439.

Stern, D.I. 2010. Energy quality. *Ecological Economics* 69 (7): 1471–1478.

Stern, D.I. 2011. The role of energy in economic growth. *Annals of the New York Academy of Sciences* 1219: 26–51.

Stern, D.I. 2012a. Interfuel substitution: A meta-analysis. *Journal of Economic Surveys* 26: 307–331.

Stern, D.I. 2012b. Modeling international trends in energy efficiency. *Energy Economics* 34: 2200–2208.

Stern, D.I., and A. Kander. 2012. The role of energy in the industrial revolution and modern economic growth. *Energy Journal* 33 (3): 125–152.

Stiglitz, J.E. 1974a. Growth with exhaustible natural resources: The competitive economy. *Review of Economic Studies* 41: 139–152.

Stiglitz, J.E. 1974b. Growth with exhaustible natural resources: Efficient and optimal growth paths. *Review of Economic Studies* 41: 123–138.

Sue Wing, I. 2008. Explaining the declining energy intensity of the U.S. economy. *Resource and Energy Economics* 30: 21–49.

Toman, M.A., and B. Jemelkova. 2003. Energy and economic development: An assessment of the state of knowledge. *Energy Journal* 24 (4): 93–112.

Turner, K. 2009. Negative rebound and disinvestment effects in response to an improvement in energy efficiency in the UK economy. *Energy Economics* 31: 648–666.

Turner, K. 2013. 'Rebound' effects from increased energy efficiency: A time to pause and reflect. *Energy Journal* 34 (4): 25–43.

Turner, K., and N. Hanley. 2011. Energy efficiency, rebound effects and the Environmental Kuznets Curve. *Energy Economics* 33: 722–741.

Ukidwe, N.U., and B.R. Bakshi. 2007. Industrial and ecological cumulative exergy consumption of the United States via the 1997 input–output benchmark model. *Energy* 32: 1560–1592.

van Benthem, A.A. *2015*. *Energy leapfrogging*. *Journal of the Association of Environmental and Resource Economists 2* (*1*): *93–132*.

Wang, C. 2011. Sources of energy productivity growth and its distribution dynamics in China. *Resource and Energy Economics* 33: 279–292.

Warr, B., and R.U. Ayres. 2010. Evidence of causality between the quantity and quality of energy consumption and economic growth. *Energy* 35: 1688–1693.

Warr, B., R.U. Ayres, N. Eisenmenger, F. Krausmann, and H. Schandl. 2010. Energy use and economic development: A comparative analysis of useful work supply in Austria, Japan, the United Kingdom and the US during 100 years of economic growth. *Ecological Economics* 69: 1904–1917.

Wei, C., J. Ni, and M. Shen. 2009. Empirical analysis of provincial energy efficiency in China. *China & World Economy* 17 (5): 88–103.

Wrigley, E. Anthony. *2010*. *Energy and the English industrial revolution*. *Cambridge*: *Cambridge University Press*.

Yu, E.S.H., and J.C. Jin. 1992. Cointegration tests of energy consumption, income, and employment. *Resources and Energy* 14: 259–266.

Enforcement

Kevin Roberts

Enforcement, with its usual connotations of agents being compelled to behave in ways that are at variance with how they would like to act, seems a long way removed from the conventional neoclassical approach of laissez-faire inherent in a decentralized economic system. If an enforcement mechanism is viewed as a method by which rule-breaking may be discouraged then laissez-faire attempts to be a set of rules that are self-enforceable so that no formal enforcement mechanism is required (see Stigler 1970 for a useful discussion of enforcement). But it is clear that decentralizability is not, in itself, enough. For instance, the fact that tax evasion is kept in check only by legal pressure shows how the addition of a tax system to a decentralized competitive structure may destroy self-enforceability. The possibility of gain by the exercise of monopoly power shows how the rule of price-taking behaviour is not immune to the problems of enforcement.

Except when self-enforceability holds, so that agents wish to follow the given set of rules – the rules are incentive compatible – an enforcement mechanism will be necessary to prevent rule-breaking. A formal enforcement mechanism has two components. First, there must be a method of monitoring agents so that it is possible to observe, perhaps imperfectly, rule-breaking. Second, there must exist a sanction or punishment which can be imposed on rule-breakers. If perfect observability is costless and there is no limit to the disutility that a punishment can impose then there is no difficulty in enforcing a rule and, importantly, the enforcement procedure never needs to be exercised. Thus, the existence of an enforcement procedure alters the incentives to obey rules and can be used as a component of the incentive structure within an economic system. It is clear that this is in part the role of a legal system.

To examine enforcement procedures in greater detail, first consider the monitoring mechanism. If monitoring is impossible then agents will abide with a rule only if it is self-enforceable. If monitoring is costly then there may be a decision to be taken with regard to the best monitoring mechanism. Two dimensions of choice seem natural. First, there is the quality of the monitoring mechanism. After monitoring, an inference will be made with regard to whether a rule has been broken. The better the quality of the monitoring mechanism, the more accurate will be this inference. Accuracy will involve both a low probability of inferring rule-breaking when it has not occurred and a low probability of inferring rule-compliance when rule-breaking has occurred.

The second component of the monitoring mechanism is the intensity with which it is applied. The same agent may be monitored several times to improve the accuracy of the test; agents may be monitored on a random basis so that any one agent will be monitored with a probability of less than unity. The overall mechanism is likely to be more effective if agents must decide upon their behaviour with regard to rule-compliance before they know the intensity with which they will be monitored.

A high punishment level will deter rule-breaking more than a low punishment level. To see what an optimal enforcement mechanism may look like, consider a planner who wishes to deter rule-breaking (which, for instance, imposes significant negative externalities on other agents). Assume that the planner wishes to maximize the expected utility of agents whilst deterring rule-breaking – rule-breaking is assumed to be privately optimal. If a is the accuracy of the test used then let $p(a)$ be the probability of a rule-breaking inference if rule-breaking has not occurred. Similarly, let $q(a)$ be the probability of such an inference if rule-breaking has occurred.

Thus, $p(a) < q(a)$, $p(a)$ is falling in a and $q(a)$ is rising in a. Let U be the utility that an agent gets from rule-compliance before the introduction of an enforcement mechanism and V be the utility that would result from rule-breaking. Then if m is the probability of being monitored and f is the punishment cost then, taking the

simplest specification of linear utility, the expected utility of an agent who complies with the rule is given by

$$U - mp(a)f - am \tag{1}$$

where am is taken to be the cost of operating the enforcement mechanism, the money being raised through taxation, say. The punishment f is assumed to be a deadweight loss rather than a monetary fine which could be used to finance the monitoring mechanism. An agent compares (1) with the expected utility from rule-breaking:

$$V - mq(a)f - am \tag{2}$$

so that rule-breaking can be deterred if

$$mf[q(a) - p(a)] \geq V - U. \tag{3}$$

If the planner wishes to maximize (1) subject to (3) then the first point to notice is that (1) can always be increased while still satisfying (3) if mf is held constant but m is reduced. Thus, it is desirable to increase the punishment whilst reducing the rate of monitoring and it is optimal to punish as severely as possible (this argument is credited to Becker 1968). Once the worst punishment is chosen, the optimal monitoring mechanism satisfies

$$\frac{fp' + 1}{fp + a} = \frac{q' - p'}{q - p} \tag{4}$$

which gives rise to the comparative statics results that would be expected. If f can be chosen without limit (punishment can reduce utility to minus infinity) then m will be set arbitrarily small and, keeping mf fixed, (1) can be increased by increasing a if $p'(a)$ is strictly negative. If perfect accuracy is attainable ($p(a) = 0$ for some a) then the expected utility of the agent will be U – the first-best is achievable – and the enforcement mechanism will take the form of an infinitesimally small amount of monitoring using a very accurate procedure and infinitely large punishments being imposed on rule-breakers. This strong result depends upon the strong assumptions that have been imposed on the model. However, Nalebuff and Scharfstein (1985) obtain a similar result in a richer setting than that of the above model. For a model where finite levels of punishment are optimal because of the risk aversion of agents, see Polinsky and Shavell (1979).

The structure outlined above has the property that the enforcement mechanism is operated by a planner. One reason for assuming this is that it is unnecessary to consider the planner's incentives to operate the mechanism. In some situations, it is necessary for a group of agents to operate an enforcement mechanism that will be imposed upon themselves. The classic example of this is the operation of a cartel (Stigler 1964). A major problem introduced by not having an external agency is that costs may be imposed on firms if they choose to punish a firm that breaks the cartel's rules. An example of this is when a firm is punished by all firms entering into a price-cutting war. To provide an incentive to punish, it may be necessary to have a mechanism to enforce the operation of the original enforcement mechanism, and so on. For an insightful analysis of this problem, see Abreu (1986). Considerations of this sort show the potential richness of the structure of enforcement mechanisms in general.

See Also

- Cartels
- Cooperative Equilibrium
- Cooperative Games
- Game Theory
- Industrial Organization

References

Abreu, D. 1986. Extremal equilibria of oligopolistic supergames. *Journal of Economic Theory* 39(1): 191–225.

Becker, G. 1968. *Crime and punishment: an economic approach*, vol March-April, 169–217. *Journal of Political Economy*.

Nalebuff, B. and Scharfstein, D. 1985. Self-selection and testing. Forthcoming in *Review of Economic Studies*.

Polinsky, A., and S. Shavell. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69(5): 880–891.

Stigler, G. 1964. A theory of oligopoly. *Journal of Political Economy* 72(1): 44–61.

Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78(3): 526–535.

Engel Curve

Arthur Lewbel and H. S. Houthakker

Abstract

An Engel curve describes how a consumer's purchases of a good like food varies as the consumer's total resources such as income or total expenditures vary. Engel curves may also depend on demographic variables and other consumer characteristics. A good's Engel curve determines its income elasticity, and hence whether the good is an inferior, normal, or luxury good. Empirical Engel curves are close to linear for some goods, and highly nonlinear for others. Engel curves are used for equivalence scale calculations and related welfare comparisons, and determine properties of demand systems such as aggregability and rank.

Keywords

Aggregation; Consumers' expenditure; Consumer demand; Demand equations; Engel curves; Engel equivalence scales; Engel's law; Law of one price; Nonparametric methods; Rothbarth scales; Separability; Utility theory; Working, H.; Working–Leser model

JEL Classifications

D12

An Engel curve is the function describing how a consumer's expenditures on some good or service relate to the consumer's total resources, with prices fixed, so $q_i = g_i(y, z)$, where q_i is the quantity consumed of good i, y is income, wealth, or total expenditures on goods and services, and z is a vector of other characteristics of the consumer, such as age and household composition. Usually y is taken to be total expenditures, to separate the problem of allocating total consumption to various goods from the decision of how much to save or dissave out of current income. Engel curves are frequently expressed in the budget share form $w_i = h_i[\log(y), z]$ where w_i is the fraction of y that is spent buying good i. The goods are typically aggregate commodities such as total food, clothing or transportation, consumed over some weeks or months, rather than discrete purchases. Engel curves can be defined as Marshallian demand functions, with the prices of all goods fixed.

The term 'Engel curve' is also used to describe the empirical dependence of q_i on y, z in a population of consumers sampled in one time and place. This empirical or statistical Engel curve coincides with the above theoretical Engel curve definition if the law of one price holds (all sampled consumers paying the same prices for all goods), and if all consumers have the same preferences after conditioning on z and possibly on some well-behaved error terms. Since these conditions rarely hold, it is important in practice to distinguish between these two definitions.

Using data from Belgian surveys of working class families, Ernst Engel (1857, 1895) studied how households' expenditures on food vary with income. He found that food expenditures are an increasing function of income and of family size, but that food budget shares decrease with income. This relationship of food consumption to income, known as Engel's law, has since been found to hold in most economies and time periods, often with the function h_i for food i close to linear in $\log(y)$.

Engel curves can be used to calculate a good's income elasticity, which is roughly the percentage change in q_i that results from a one per cent change in y, or formally $\partial \log g_i(y, z)/\partial \log(y)$. Goods with income elasticities below zero, between zero and 1, and above 1 are called inferior goods, necessities and luxuries respectively, so by these definitions what Engel found is that food is a necessity. Elasticities can themselves vary with income, so a good that is a necessity for the rich can be a luxury for the poor.

Some empirical studies followed Engel (1895), such as Ogburn (1919), but Allen and Bowley (1935) firmly connected their work to utility theory. They estimated linear Engel curves $q_i = a_i + b_i y$ on data-sets from a range of countries, and found that the resulting errors in these models were sometimes quite large, which they interpreted as indicating considerable

heterogeneity in tastes across consumers. Working (1943) proposed the linear budget share specification $w_i = a_i + b_i \log(y)$, which is known as the Working–Leser model, since Leser (1963) found this functional form to fit better than some alternatives. However, Leser obtained still better fits with what would now be called a rank-three model, namely, $w_i = a_i + b_i \log(y) + c_i y^{-1}$, and in a similar, earlier, comparative statistical analysis Prais and Houthakker (1955) found $q_i = a_i + b_i \log(y)$ to fit best. More recent work documents sometimes considerable nonlinearity in Engel curves. Motivated by this nonlinearity, one of the earlier empirical applications of nonparametric regression methods in econometrics was kernel estimation of Engel curves. Examples include Bierens and Pott-Buter (1990), Lewbel (1991), and Härdle and Jerison (1991). More recent studies that control for complications like measurement error and other covariates Z, including Hausman et al. (1995) and Banks et al. (1997), find Engel curves for some goods are close to Working–Leser, while others display considerable curvature, including quadratics or S shapes. Even Allen and Bowley (1935, p. 123) noted 'there is a good fit, allowance being made for observation and sampling errors,. . ., to a linear expenditure relation and occasionally to a parabolic relation'.

Other variables z also help explain cross-section variation in demand. Commonly used covariates include the number, ages and gender of family members, location measures, race and ethnicity, seasonal effects, and labour market status. Variables indicating ownership of a home, a car or other large durables can also have considerable explanatory power, though these are themselves consumption decisions.

Engel's original work showed the relevance of family size, and later studies confirm that larger families typically have larger budget shares of necessities than smaller families at the same income level. Adult equivalence scales model the dependence of utility functions on family size, and use this dependence to compare welfare across households, assuming that a large family with a high income is as well off as a smaller family with a lower income if both families have demands that are similar in some way, such as equal food budget shares or equal expenditures on adult goods such as alcohol. The ratio of total expenditures needed to equate food budget shares across households are known as Engel equivalence scales, while the ratio that equates expenditures on adult goods are called Rothbarth scales (Rothbarth 1943).

Shape invariance assumes that budget share Engel curves for one type of consumer, such as a household with children, is a linear transformation of the budget-share Engel curves for other types of consumers, such as households without children. Shape invariance is necessary for constructing what are known as exact or independent of base equivalence scales, and has been found to at least approximately hold in some data-sets. See Lewbel (1989), Blackorby and Donaldson (1991), Gozalo (1997), Pendakur (1999), and Blundell et al. (2003).

The level of aggregation across goods affects Engel curve estimates. Demand for a narrowly defined good like apples varies erratically across consumers and over time, while Engel curves based on broad aggregates like food are affected by variation in the mix of goods purchased. The aggregate necessity food could include inferior goods like cabbage and luxuries like caviar, which may have very different Engel curve shapes.

Other empirical Engel curve complications include unobserved variations in the quality of goods purchased, and violations of the law of one price. When price or quality variation is unobserved, their effects may correlate with, and so be erroneously attributed to, y or z. Examples of such correlations could include the wealthy systematically favouring higher quality goods, and the poor facing higher prices than other consumers because they cannot afford to travel to discount stores.

Assume a consumer (household) h determines demands q_{hi} facing prices p_i for each good i by maximizing a well-behaved utility function over goods (which could depend on z_h), subject to a budget constraint $\sum_i p_i q_{hi} \leq y_h$. This yields Marshallian demand functions $q_{hi} = G_{hi}(P, y_h, z_h)$, with Engel curves given by these functions with the price vector p fixed. Utility functions that yield Engel curves of the form $q_{hi} = b_i(z) y_h$ are called homothetic, and $q_{hi} = a_i(z) + b_i(z) y_h$ are quasihomothetic. Many theoretical results

regarding two-stage budgeting and aggregation across goods require homotheticity or quasi-homotheticity, most notably Gorman (1953).

The shape of Engel curves plays an important role in the determination of macroeconomic demand relationships. For example, if we ignore z for now, suppose individual consumers h each have Engel curves of the quasihomothetic form $q_{hi} = a_{hi} + b_i y_h$. Then, letting Q_i and Y be aggregate per capita quantities and total expenditures in the population, we get $Q_i = A_i + b_i Y$ by averaging q_{hi} across consumers h. This is a representative consumer model, in the sense that the distribution of y affects aggregate demand Q_i only through its mean $E(y) = Y$. Gorman (1953) showed that only linear Engel curves have this property, though linear Engel curve aggregation dates back at least to Antonelli (1886). Gorman's linearity requirement, which does not usually hold empirically, can be relaxed given restrictions on the distribution of y; for example, Lewbel (1991) shows that $E(y \log y)/Y - \log(Y)$ is very close to constant in US data, and if it is constant then Working–Leser household Engel curves yield Working–Leser aggregate, representative consumer demands.

Exactly aggregable demands are defined by $q_i = \sum_{j=1}^{J} A_{ji}(p) c_j(y, z)$, and so have Engel curves $q_i = \sum_{j=1}^{J} a_{ji} c_j(y, z)$that are linear in the functions $cj(y, z)$. These models have the property that aggregate demands Q_i depend only on the means of $c_j(y, z)$. Utility theory imposes constraints on the functional forms of $c_j(y, z)$. Properties of exactly aggregable demands and associated Engel curves are derived in Muellbauer (1975), Jorgenson et al. (1982), and Lewbel (1990), but primarily by Gorman (1981), who proved the surprising result that utility maximization forces the matrix of Engel curve coefficients a_{ji} to have rank three or less.

Lewbel (1991) extends Gorman's rank idea to arbitrary demands, not just those in the exactly aggregable class, by defining the rank of a demand system as the dimension of the space spanned by its Engel curves. Engel curve rank can be nonparametrically tested, and has implications for utility function separability, welfare comparisons, and for aggregation across goods and across consumers. Many empirical studies find demands have rank three.

One area of current research concerns the observable implications of collective models, that is, households that determine expenditures based on bargaining among members. For example, the Engel curves of such households could violate Gorman's rank theorem, even if each member had exactly aggregable preferences. Another topic attracting current attention is the role of errors in demand models, particularly their interpretation as unobserved preference heterogeneity, random utility model parameters. This matters in part because another of Allen and Bowley's (1935) findings remains true today, namely, Engel curve and demand function models still fail to explain most of the observed variation in individual consumption behaviour.

See Also

- ▶ Aggregation (Theory)
- ▶ Consumer Expenditure
- ▶ Demand Theory
- ▶ Engel, Ernst (1821–1896)
- ▶ Engel's Law
- ▶ Equivalence Scales
- ▶ Gorman, W.M. (Terence)
- ▶ Utility

Bibliography

Allen, R., and A. Bowley. 1935. *Family expenditure: A study of its variation*. London: P.S. King and Son.

Antonelli, G. 1886. *Sulla teoria metematica della economia politica*. Pisa: Nella Tipografia del Fochetto. Translated as 'On the mathematical theory of political economy'. In *Preferences, utility and demand,* ed. J. Chipman et al. New York: Harcourt Brace Jovanovich, 1971.

Banks, J., R. Blundell, and A. Lewbel. 1997. Quadratic Engel curves and consumer demand. *The Review of Economics and Statistics* 79: 527–539.

Bierens, H., and H. Pott-Buter. 1990. Specification of household expenditure functions and equivalence scales by nonparametric regression. *Econometric Reviews* 9: 123–210.

Blackorby, C., and D. Donaldson. 1991. Adult-equivalence scales, interpersonal comparisons of well-being, and applied welfare economics. In *Interpersonal*

comparisons of well-being, ed. J. Elster and J. Roemer. Cambridge: Cambridge University Press.
Blundell, R., M. Browning, and I. Crawford. 2003. Nonparametric Engel curves and revealed preference. *Econometrica* 71: 205–240.
Engel, E. 1857. Die Productions- und Consumptionsverhaeltnisse des Koenigsreichs Sachsen. *Zeitschrift des Statistischen Bureaus des Koniglich Sachsischen Ministeriums des Inneren, No. 8 und 9*. Reprinted in the Appendix of Engel (1895).
Engel, E. 1895. Die Lebenskosten Belgischer Arbeiter-Familien Fruher and jetzt. *International Statistical Institute Bulletin* 9: 1–74.
Gorman, W. 1953. Community preference fields. *Econometrica* 21: 63–80.
Gorman, W. 1981. Some Engel curves. In *Essays in the theory and measurement of consumer behaviour in honor of Sir Richard Stone*, ed. A. Deaton. Cambridge: Cambridge University Press.
Gozalo, P. 1997. Nonparametric bootstrap analysis with applications to demographic effects in demand functions. *Journal of Econometrics* 81: 357–393.
Härdle, W., and M. Jerison. 1991. Cross section Engel curves over time. *Recherches Economiques de Louvain* 57: 391–431.
Hausman, J., W. Newey, and J. Powell. 1995. Nonlinear errors in variables: Estimation of some Engel curves. *Journal of Econometrics* 65: 205–253.
Jorgenson, D., L. Lau, and T. Stoker. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in econometrics*, ed. R. Basman and G. Rhodes. Greenwich: JAI Press.
Leser, C. 1963. Forms of Engel functions. *Econometrica* 31: 694–703.
Lewbel, A. 1989. Household equivalence scales and welfare comparisons. *Journal of Public Economics* 39: 377–391.
Lewbel, A. 1990. Full rank demand systems. *International Economic Review* 31: 289–300.
Lewbel, A. 1991. The rank of demand systems: Theory and nonparametric estimation. *Econometrica* 59: 711–730.
Muellbauer, J. 1975. Aggregation, income distribution, and consumer demand. *Review of Economic Studies* 62: 269–283.
Ogburn, W. 1919. Analysis of the standard of living in the District of Columbia in 1916. *Journal of the American Statistical Association* 16: 374–389.
Pendakur, K. 1999. Estimates and tests of base-independent equivalence scales. *Journal of Econometrics* 88: 1–40.
Prais, S., and H. Houthakker. 1955. *The analysis of family budgets*. 2nd ed. Cambridge: Cambridge University Press. 1971.
Rothbarth, E. 1943. Note on a method of determining equivalent income for families of different composition. In *Appendix 4 in War-time pattern of saving and spending*, ed. C. Madge. Cambridge: Cambridge University Press.
Working, H. 1943. Statistical laws of family expenditures. *Journal of the American Statistical Association* 38: 43–56.

Engel, Ernst (1821–1896)

H.S. Houthakker

Keywords

Agricultural economics; Engel curve; Engel, E.; Engel's Law; Household surveys; International Statistical Institute

JEL Classifications

B31

Born in Dresden, Engel was a German statistician best known for the discovery of the Engel curve and of Engel's Law. In his early years he was associated with the French sociologist Frédéric Le Play, whose interest in the family led him to conduct household surveys. The expenditure data collected in these surveys convinced Engel that there was a relation between a household's income and the allocation of its expenditures between food and other items. This was one of the first functional relations ever established quantitatively in economics. Furthermore, he observed that households with higher incomes tended to spend more on food than poorer households, but that the share of food expenditures in the total budget tended to vary inversely with income. From this empirical regularity he went on to infer that in the course of economic development agriculture would decline relative to other sectors of the economy (Engel, 1857). From 1860 to 1882 Engel was director of the Prussian statistical bureau in Berlin, in which capacity he did much to expand and strengthen official statistics. His resignation resulted from his opposition to Bismarck's protectionist policies. In his own research he dealt particularly with the value of human life (Engel, 1877), which he approached from the cost side. He also investigated the influence of price on demand. His influence on official statistics extended well beyond Germany, and in 1885 he was among the founders of the

International Statistical Institute. He died in Radebeul in 1896.

Selected Works

1857. Die Productions- und Consumptionsverhaeltnisse des Koenigsreichs Sachsen. Reprinted with Engel (1895), *Anlage* I: 1–54.
1877. *Der Kostenwerth des Menschen*. Berlin.
1895. Die Lebenskosten Belgisher Arbeiter-Familien fruether und jetzt. Reprinted in *International Statistical Institute Bulletin* 9: 1–124.

Engel's Law

Martin Browning

Keywords
Agriculture and economic development; Equivalence scales; Engel's law

JEL Classifications
D12

Engel's law states that food is not a luxury. This is one of the earliest empirical regularities in economics and also one of the most robust. The widespread finding is that regressions of food expenditures, quantities or budget shares on income or total expenditure and other variables such as prices, demographics and regional dummies uniformly imply that the income elasticity of food is less than 1 (and greater than zero). For example, time series from individual countries, cross-sections within countries and cross-country analyses all find the same qualitative empirical finding.

This correlation seems to have been highlighted for a number of reasons. First, food is an important component of household budgets everywhere so that it is intrinsically of interest. Second, the finding suggests that over the long run countries experiencing significant growth will find that agriculture provides an increasingly unimportant part of national income. This argues against balanced growth in long- run development. Third, we do not observe such a consistent pattern for any other wide commodity grouping such as clothing or durables. Finally, the fact that the food budget share is a decreasing function of the material standard of living (if other factors are held constant) suggested at one time that it can be used as an indicator of the latter. In particular, iso-prop ('same proportion') methods have been used to compute adult equivalence scales by finding the level of income that would equate the food budget share across different demographic groups. The conditions under which the iso-prop method is valid are very strong – essentially, extra people in the household have to make the household behave as though it is poorer and should not cause any change in the structure of demands above this – and such methods have fallen out of favour (see Deaton and Muellbauer 1986, for discussion and references).

Despite the venerability of the literature on Engel's law, the inferences that can be drawn from it are limited. For example, the cross-section finding is consistent with all households having a decreasing relationship so that increasing the income of a household will lead to a decrease in the food budget share. On the other hand, the correlation might be completely spurious if it is due to poorer households having a higher 'taste' for food. In this case the apparent dependence is simply due to heterogeneity in tastes, which is correlated with income. The fact that studies using aggregate time series-data find different elasticities from those found in cross-section data from the same country and time period suggests that the empirical finding is a combination of both causes. The paucity of panel data with full expenditure information makes any inference hazardous. Thus Engel's law remains what it has always been: a very robust but unsurprising partial correlation with many alternative interpretations.

See Also

- ▶ Engel Curve
- ▶ Engel, Ernst (1821–1896)
- ▶ Equivalence Scales

Bibliography

Deaton, A., and J. Muellbauer. 1986. On measuring child costs: With applications to poor countries. *Journal of Political Economy* 94: 720–744.

Engels, Friedrich (1820–1895)

Gareth Stedman Jones

Keywords

Cost of production theory of value; Engels, F.; Kautsky, K.; Labour theory of value; Marx, K. H.; Marxian transformation problem; Peasants; Rent; Subjective theory of value

JEL Classifications

B31

Born in Barmen, the eldest son of a textile manufacturer in Westphalia, Engels was trained for a merchant's profession. From school onwards, however, he developed radical literary ambitions which eventually brought him into contact with the Young Hegelian circle in Berlin in 1841. In 1842, Engels left for England to work in his father's Manchester firm. Already converted by Moses Hess to a belief in 'communism' and the imminence of an English social revolution, he used his two-year stay to study the conditions which would bring it about. From this visit came two works which were to make an important contribution to the formation of Marxian socialism: *Outlines of a Critique of Political Economy* (generally called the *Umrisse)* published in 1844, and *The Condition of the Working Class in England,* published in Leipzig in 1845.

Returning home via Paris in 1844, Engels had his first serious meeting with Marx. Their lifelong collaboration dated from this point with an agreement to produce a joint work (*The Holy Family*), setting out their positions against other tendencies within Young Hegelianism. This was followed by a second unfinished joint enterprise *(The German Ideology,* 1845–6), where their materialist conception of history was expounded systematically for the first time.

Between 1845 and 1848, Engels was engaged in political work among German communist groups in Paris and Brussels. In the 1848 revolution itself, he took a full part, first as a collaborator of Marx on the *Neue Rheinische Zeitung* and subsequently in the last phase of armed resistance to counter-revolution in the summer of 1849.

In 1850, Engels returned once more to Manchester to work for his father's firm and remained there until he retired in 1870. During this period, in addition to numerous journalistic contributions, including attempts to publicize Marx's *Critique of Political Economy* (1859) and *Capital,* volume 1 (1867, second edition 1873), he first developed his interest in the relationship between historical materialism and the natural sciences. These writings were posthumously published as *The Dialectics of Nature* (1925). In 1870 Engels moved to London.

As Marx's health declined, Engels took over most of his political work in the last years of the First International (1864–72) and took increasing responsibility for corresponding with the newly founded German Social Democratic Party and other infant socialist parties. Engels's most important work during this period was his polemic against the positivist German socialist, Eugen Dühring. The *Anti-Dühring* (1877) was the first comprehensive exposition of a Marxian socialism in the realms of philosophy, history and political economy. The success of this work, and in particular of extracts from it like *Socialism, Utopian and Scientific,* represented the decisive turning point in the international diffusion of Marxism and shaped its understanding as a theory in the period before 1914.

In his last years after Marx's death in 1883, Engels devoted most of his time to the editing and

publishing of the remaining volumes of *Capital* from Marx's manuscripts. volume 2 appeared in 1885, volume 3 in 1894, a year before his death. Engels had also hoped to prepare the final volume dealing with the history of political economy. But the difficulty of deciphering Marx's handwriting, his own failing eyesight and the formidable editorial problems encountered in constructing Volumes 2 and 3, induced him to hand over this task to Karl Kautsky, who subsequently published it under the title *Theories of Surplus Value.*

Engels's work was of importance, both in the construction and interpretation of Marxian economic theory and in the laying down of important guidelines in the subsequent development of Marxist economic policy.

In the realm of theory, his contribution is of particular significance in three respects.

First, and of real importance in the formation of a distinctively Marxian stance towards political economy was Engels's *Outlines of a Critique of Political Economy* (the *Umrisse),* published in 1844. In 1859 in his own *Critique of Political Economy,* Marx acknowledged this sketch as 'brilliant', and its impact is discernible in Marx's 1844 writings. The *Umrisse* represented the first systematic confrontation between the 'communist' strand of Young Hegelianism and political economy. The communist aspiration was expressed in Feuerbachian language, while the mode of analysis was Hegelian. But, as has recently been demonstrated (Claeys 1984), the content of Engels's critique was first and foremost a product of his early stay in Manchester. For, apart from some indebtedness to Proudhon's *What is Property?* (1841), the main source of Engels's essay was John Watts, *The Facts and Fictions of Political Economy* (1842), a resumé of the Owenite case against the propositions of political economy. At this stage, Engels's own acquaintance with the work of political economists seems to have been mainly at second-hand.

The *Umrisse* was an attempt to demonstrate that all the categories of political economy presupposed competition which in turn presupposed private property. He began with an analysis of value, which juxtaposed a 'subjective' conception of value as utility ascribed to Say with an 'objective' conception as cost of production attributed to Ricardo and McCulloch. Reconciling these two definitions in Hegelian fashion, Engels defined value as the relation of production costs to utility. This was the equitable basis of exchange, but one impossible to implement on the basis of competition which was responsive to market demand rather than social need. (Engels still adhered to this definition of value 30 years later in the *Anti-Dühring.* Discussing the disappearance of the 'law of value' with the end of commodity production, he wrote:

> As long ago as 1844, I stated that the above mentioned balancing of useful effects and expenditure of labour would be all that would be left, in a communist society, of the concept of value as it appears in politicaleconomy . . . The scientific justification for this statement, however,. . . was only made possible by Marx's *Capital.* (Engels 1877, pp. 367–8)

This shows how much greater continuity of thought there was between the young and the old Engels than is normally imagined.)

He next analysed rent, counterposing a Ricardian notion of differential productivity to one attributed to Smith and T.P. Thompson based upon competition. Interestingly, in this analysis Engels differed both from Watts and Proudhon, in denying the radical form of the labour theory – the right to the whole product of labour – both by citing the case of the need to support children and in querying the possibility of calculating the share of labour in the product.

Finally, after an attack on the Malthusian population theory, which closely followed Alison and Watts, Engels attacked competition itself, both because it provided no mechanism of reconciling general and individual interest, and because it was argued to be self-contradictory. Competition based on self-interest bred monopoly. Competition as an immanent law of private property led to polarization and the centralization of property. Thus private property under competition is self-consuming.

What particularly impressed Marx was the argument that all the categories of political economy were tied to the assumption of competition based on private property. This, for him,

represented an important advance over Proudhon whose notion of equal wage would lead to a society conceived as 'abstract capitalist' and whose conception of labour right presupposed private property. Proudhon had not seen that labour was the essence of private property. His critique was of 'political economy from the standpoint of political economy'. He had not 'considered the further creations of private property, e.g. wages, trade, value, price, money etc. as forms of private property in themselves' (Engels and Marx 1844b). The *Umrisse* suggested a new means of underpinning the Marxian ambition to transcend the categorical world of political economy and private property altogether. Moreover, by representing competition as a law which would produce its opposite, monopoly, the elimination of private property and revolution, Engels preceded Marx in positing the 'free trade system' as a process moving towards self-destruction through the operation of laws immanent within it.

These conclusions were amplified in Engels's other major work of this period, *The Condition of the Working Class in England.* Here, the law of competition by engendering 'the industrial revolution' had created a revolutionary new force, the working class. The single thread underlying the development of the working class movement had been the attempt to overcome competition. Such an analysis prefigured the famous statement in the *Communist Manifesto* that the capitalists were begetting their own gravediggers (Stedman Jones 1977).

Between the mid-1840s and the mid-1870s, Engels played no discernible part in the elaboration of *Capital* beyond supplying Marx with practical business information. His vital contributions to the prehistory of the theory were forgotten and it was only in his better-known role as interpreter and publicist of Marx's work that his writings received widespread attention. During the Second International period, these writings attained almost canonical status, but in the 20th century they generally provided a polemical target for all those attempting to re-theorize Marx in the light of the publication of his early writings.

In the realm of political economy more narrowly conceived, Engels helped to set up the 'transformation' debate by his dramatization of Marx's switch from value to production price in his introductions to Volumes 2 and 3 of *Capital.* Engels's own contribution to this debate in his last published article in *Neue Zeit* in 1895 (now published as 'Supplement and Addendum' to volume 3 of *Capital)* was to argue that the shift from value to production price was not merely a logical development entailed by the enlargement of the scope of investigation to include circulation and the 'process of capitalist production as a whole', but also reflected a real historical transition from the stage of simple commodity production to that of capitalism proper. 'The Marxian law of value has a universal economic validity for an era lasting from the beginning of the exchange that transforms products into commodities down to the fifteenth century of our epoch' (Marx 1894, p. 1037).

Leaving aside the empirical question whether during the pre-capitalist era commodities were exchanged in accordance with the amount of labour embodied in them, commentators as diverse as Bernstein and Rubin have objected that this makes no sense in terms of Marx's theory, since during this epoch there existed 'no mechanism of the general equalisation of different individual labour expenditures in separate economic units on the market' and that consequently it was not appropriate to speak of 'abstract and socially necessary labour which is the basis of the theory of value' (Rubin 1928, p. 254). They have further objected, appealing to Marx's 1857 'Introduction to the Critique of Political Economy', that there is no necessary connection between the logical and historical sequence of concepts, and that the order of appearance of concepts in *Capital* is determined simply by the logical place they occupy in an exposition of the theory of the capitalist mode of production.

Engels could certainly claim explicit textual support from volume 3 for his historical interpretation of value ('It is also quite apposite to view the value of commodities not only as theoretically prior to the prices of production, but also as historically prior to them. This applies to those conditions in which the means of production belong to the worker. . .'; Marx 1894, p. 277). It should also be stressed that there was nothing new in

Engels's representation of the character of Marx's theory. Back in 1859, in a review of Marx's *Critique of Political Economy,* Engels stated, 'Marx was, and is, the only one who could undertake the work of extracting from the Hegelian Logic the kernel which comprised Hegel's real discoveries . . . and to construct the dialectical method divested of its idealistic trappings'; and in characterizing that method as a form of identity between logical and historical progression, he continued, 'the chain of thought must begin with the same thing that this history begins with, and its further course will be nothing but the mirror image of the historical course in abstract and theoretically consistent form . . . ' (Engels 1859). It is implausible to suppose that Marx at this time should have sanctioned a fundamental distortion of his method and it is suggestive that he himself, describing his relationship to Hegel, should have endorsed the metaphor of discovering 'the rational kernel in the mystical shell' in his 1873 Postface to the second edition of *Capital,* volume 1 (Marx 1873, p. 103). Perhaps the real difficulty lies not in Engels but in Marx himself. It may be, as Louis Althusser has claimed, that Marx did not find a suitable language in which to characterize the distinctiveness of his approach, or it may be more simply that Marx remained ambivalent about how to characterize the theory. In any event, it is not difficult to establish disjunctions between the way he proceeds and the descriptions he gives of his procedures. Engels stuck fairly closely to Marx's descriptions of his procedures and can hardly be reproached for taking Marx at his word.

The problem of Engels's role as an interpreter of Marx's theory debouches onto a third and potentially yet more contentious aspect of Engels's legacy, his role as editor of *Capital,* Volumes 2 and 3. Engels's work was not confined to the transcription of Marx's illegible handwriting. He had to make active editorial choices. The published versions of these volumes contain over 1300 pages, but the original manuscripts amount to almost twice as many. For volume 2, for instance, Marx had composed eight versions of his treatment of the process of circulation, from which Engels made a collation. In the absence of an independent transcription and publication of the manuscripts, from which Engels worked, it is impossible to assess whether the emphasis and meaning of the published volumes differ in any significant way from the original. What seems clear, is that in his cautious desire to reproduce as much of the original material as possible, Engels produced a much bulkier and more repetitive version than Marx originally intended. Marx, it seems, always hoped that *Capital* should consist of two volumes and a further volume on the history of political economy (Rubel 1968; Levine 1984). From a detailed comparison of volume 2, Part 1, with the original manuscripts, it appears that Engels also occasionally committed inaccuracies in the citation of the manuscripts he had used (Levine 1984). Much more doubtful, given all we know of Engels's caution as an editor, is the further suggestion that Engels's editing procedures may have shifted the meaning of the text in ways that lent support to a 'collapse theory' of capitalism *(Zusammenbruchstheorie)* (Levine 1984). Apart from the smallness of the sample and Engels's own reservations about such a theory, the fact is that proponents of such a position already had sufficient ammunition from *Capital*, volume 1. Moreover, it simply begs the question whether Marx's attitude to the collapse of capitalism was any more or less apocalyptic than that of Engels.

This discussion by no means exhausts Engels's importance in the history of economic theory or policy. A fuller treatment would have to discuss his analysis of the 'peasant question' which included the important prescription that collectivization must be by example rather than force, his definition of political economy in the *Anti-Dühring,* his interpolations in *Capital,* volume 3, on banks, the stock exchange and cartels which set the agenda for the early 20th-century discussion of finance capital, his various writings on the relationship between the state and economic forces and his later surveys of English developments since 1844 which prepared the way for later Marxist theories of labour aristocracy. These are only some of the more salient examples.

Finally, at a time when it seems that the technical debate on value seems to have reached a moment of exhaustion, it is perhaps worth going

back to Engels if only to remind us of the anti-economic purpose underlying Marx's attempt to construct a theory of value in the first place.

Selected Works

1844a. Outlines of a critique of political economy. In Karl Marx and Frederick Engels, *Collected works* [MECW], vol. 3. London: Lawrence & Wishart, 1975.
1844b. (With K. Marx). *The holy family.* In MECW, vol. 4.
1845. *The condition of the working class in England*. In MECW, vol. 4.
1845–6. (With K. Marx). *The German ideology.* London: Lawrence & Wishart, 1987.
1859. Karl Marx. A contribution to the critique of political economy. *Das Volk*, Nos. 14 and 16, 6 and 20 August.
1877. *Anti-Dühring.* Moscow: Foreign Languages Publishing House, 1954.
1894. *The peasant question in France and Germany.* In Karl Marx and Frederick Engels, *Selected works*, vol. 3. Moscow: Progress Publishers, 1970.
1925. *The dialectics of nature.* Moscow: Foreign Languages Publishing House, 1954.
1938. *Engels on capital.* London: Lawrence & Wishart.

Bibliography

Claeys, G. 1984. *Engels'* outlines of a critique of political economy (1843) and the origins of the Marxist critique of capitalism. *History of Political Economy* 16: 207–232.
Levine, N. 1984. *Dialogue within dialectics*. London: Allen & Unwin.
Marx, K. 1859. A contribution to the critique of political economy. In *MECW*, vol. 16. New York: International Publishers.
Marx, K. 1873. *Capital.* Vol. 1, 2nd ed. Harmondsworth: Penguin, 1976.
Marx, K. 1894. *Capital.* Vol. 3. Harmondsworth: Penguin, 1981.
Rubel, M., ed. 1968. *Karl Marx, Oeuvres*. Vol. 2. Paris: Gallimard.
Rubin, I. 1928. *Essays on Marx's theory of value*, 1972. Detroit: Black & Red.
Stedman Jones, G. 1977. Engels and the history of Marxism. In *The history of Marxism,* ed. E.J. Hobsbawm. Hassocks: Harvester, 1983.

Bibliographic Addendum

Carver, T. 1990. *Friedrich Engels: His life and thought*. London: Palgrave Macmillan.
Hunley, J.D. 1991. *The life and thought of Friedrich Engels: A reinterpretation of his life and thought*. New Haven: Yale University Press, are useful in understanding Engels as an original thinker in his own right.

Engle, Robert F. (Born 1942)

Tim Bollerslev

Abstract

Robert Engle has published widely on topics ranging from urban economics to band spectrum regression, electricity demand, state-space modelling, testing, exogeneity, seasonality, option pricing, and market microstructure finance. Most notable, however, are his seminal contributions on cointegration and Auto-Regressive Conditional Heteroskedasticity (ARCH), which have revolutionized the field of time series econometrics and the practice of empirical macroeconomics and asset pricing finance, respectively. The research field of financial econometrics and corresponding developments in practical risk management and measurement also derive largely from the insights afforded by the ARCH class of models and Engle's many other research contributions since the 1980s.

Keywords

ARCH models; Asset pricing finance; Cointegration; Engle, R.; Financial econometrics; Fischer, F.; Friedman, M.; Granger Representation Theorem; Phillips curve; Risk management; Risk measurement; Rothenberg, J.; Solow, R.; Time series econometrics

JEL Classifications
B31; C1; C32; G1

Robert F. Engle was born in Syracuse, upstate New York, on 10 November 1942. Shortly thereafter his family moved to Philadelphia, and Engle graduated from high school there in 1960. He majored in physics as an undergraduate at Williams College, and went on to enrol as a Ph.D. student in physics at Cornell University. However, after one year he decided to switch to the Ph.D. programme in economics, where he wrote his thesis on temporal aggregation and the relationship between macroeconomic models estimated at different frequencies, under the direction of T.C. Liu. After graduating from Cornell in 1969, Engle was hired as an assistant professor at MIT. He moved on to University College at San Diego (UCSD) in 1975, where he was promoted to full professor in 1977 and a Chancellors' Associates Chair in 1993. He also chaired the UCSD Economics Department from 1990 to 1994. In 2000 his growing interest in financial markets prompted him to accept the Michael Armellino Professorship in Finance at the Stern School of Business at New York University, and he now lives on Manhattan with his wife of many years, Marianne, for most of the year. Together they have two grown children.

Engle has written and published extensively on a wide array of topics, ranging from urban economics to band spectrum regression, electricity demand, state-space modelling, testing, exogeneity, seasonality, option pricing, and market microstructure finance. However, he is particularly well-known for his contributions to time series econometrics and his path-breaking work on cointegration and AutoRegressive Conditional Heteroskedasticity (ARCH). The 2003 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel was explicitly awarded to Engle for 'methods of analyzing economic time series with time-varying volatility (ARCH)', a prize he shared with Clive W. J. Granger for his seminal contributions to the theory of cointegration. It is hardly an exaggeration to say that since the 1980s the concepts of cointegration and ARCH have completely revolutionized the field of time series econometrics and the practice of empirical macroeconomics and asset pricing finance, respectively. The blossoming new research field of financial econometrics and corresponding developments in practical risk management and measurement may also in large part be attributed to the insights afforded by the ARCH class of models and some of Engle's many other pioneering research contributions.

Encouraged by his senior colleagues Franklin M. Fisher, Robert Solow and Jerome Rothenberg, much of Engle's work as an assistant professor at MIT was in the area of urban economics. In fact, Engle was hired by UCSD as an urban economist, and he continued to teach, and occasionally publish in, urban economics almost up until he left San Diego in 2000. It was Clive Granger, whom Engle had first met at the 1970 World Congress of the Econometric Society in Cambridge, who persuaded Engle to move to the West Coast. Granger had himself just accepted a permanent position at UCSD in 1974 and, only a few years after Engle's arrival in 1975, Halbert White also joined the department. The ensuing two decades may rightfully be referred to as the golden age of modern time series econometrics, and UCSD, along with Yale, home of the group led by Peter Phillips, was *the* place to be. The list of visitors to the UCSD Economics Department over this period reads like a who's who in time series econometrics. Engle's hospitality and generosity with his time, as well as the many successful conferences he organized in San Diego, played a crucial role in fostering this nexus. The group was further strengthened by the arrival of James Hamilton, Graham Elliott and Allan Timmermann as additional faculty members in the early 1990s, and the Engle–Granger UCSD econometrics tradition continues to this day. Many of Engle's former Ph.D. students from that period have also gone on to successful academic careers, continuing the UCSD legacy.

Albert Einstein's famous maxim 'Everything should be made as simple as possible, but not simpler' succinctly characterizes Engle's approach to econometric modelling. Consider his early research on band spectrum regression. The static OLS regression approach routinely employed throughout economics implicitly

assumes that the identical linear relationship holds across all frequencies. Yet in many situations this is obviously a gross oversimplification. For instance, the relation between interest rates and housing starts arguably differs between the short run and the long run. Similarly, the Phillips-curve trade-off between unemployment and inflation may be primarily a business cycle phenomenon. Rather than building a fully fledged complicated dynamic model for analysing these types of temporal dependencies, the band spectrum regression approach offers a simple way of estimating separate regression coefficients, and therefore different relationships, for different frequencies. The idea of estimating different short-run and long-run regressions may also be seen as a precursor to Engle's later work on cointegration and error correction models.

The original idea of cointegration came from Granger. Nonetheless, it was the seminal joint paper by Engle and Granger (1987a) that devised the first empirical test for cointegration and formally established the link between cointegration and the error-correction type models popularized by Denis Sargan and David Hendry at the LSE during the 1960s and 1970s. More specifically, suppose that the two univariate time series yt and x_t are both non-stationary, or I(1), so that their first differences, $\Delta y_t \equiv y_t - y_{t-1}$ and $\Delta x_t \equiv x_t - x_{t-1}$, are stationary, or I(0). Most nominal macroeconomic and financial time series may be characterized in this way. Any linear combination of the two series, say $z_t = y_t - \beta x_t$, will then generally also be non-stationary. However, it is possible that z_t may actually be stationary, or I(0), in which case y_t and x_t are said to be cointegrated, with cointegrating vector $(1, -\beta)$. Indeed, many of the 'classical ratios' in macroeconomics and finance (such as consumption/income and dividends/prices) are naturally thought of as cointegrating relationships when expressed in logs. Engle and Granger showed that in this situation a satisfactory vector autoregression for the stationary bivariate process of first differences, $\{\Delta y_t, \Delta x_t\}$, must necessarily include the z_t 'error-correction' term in at least one of the two equations, the so-called Granger Representation Theorem. Intuitively, while both y_t and x_t are stochastically trending, they trend together, so that in the long run they do not stray too far apart. The inclusion of the stationary z_t term as an additional explanatory variable ensures this condition. On the other hand, if the two variables are not cointegrated z_t will be non-stationary, resulting in an unbalanced regression. Hence, empirically the null hypothesis of no cointegration may be assessed on the basis of the popular Engle–Granger cointegration test for a unit root in z_t or, if β is not known, a least-squares estimate thereof. The cointegration concept has had a profound impact on practical macroeconomic time series modelling in government and private institutions around the world. The academic literature also abounds with hundreds, if not thousands, of papers expanding upon the basic testing and modelling approach first developed by Engle and Granger. Engle's subsequent work on common features may also be seen as a natural extension of the cointegration concept.

Another more technical theme brought to the fore by Engle's research entails the powerful use of one-step-ahead prediction error decompositions and conditional Gaussian likelihoods. For instance, the beauty of his influential work on testing, including the simple-to-implement Lagrange Multiplier (LM) chi-square type test statistics constructed by multiplying the number of time series observations with the R^2 from an auxiliary regression of either unity on the vector of scores evaluated under the null hypothesis, or, alternatively, a regression of the squared residuals on the derivatives of the conditional mean, hinges directly on recursively expressing the likelihood function in terms of conditional one-step-ahead densities. Engle's pioneering contributions on dynamic factor models and Kalman filtering are similarly based on the powerful idea of representing the likelihood function in terms of successive conditional densities. Most important, however, the seminal ARCH class of models is also formulated directly in terms of one-step-ahead conditional expectations and densities.

The ARCH model (aptly named so by David Hendry) was conceived during Engle's sabbatical

visit to the LSE in 1979. Engle's interest in modelling variance dynamics was spurred by the assertion in Milton Friedman's 1976 Nobel Lecture on a trade-off between unemployment and inflationary uncertainty rather than a trade-off between unemployment and the level of inflation as stipulated by the conventional Phillips curve. The actual formulation of the first ARCH model was also influenced by Granger's ongoing work on bilinear models. At the time Granger had noted that in a non-Gaussian setting white noise series need not necessarily be unpredictable, and, in particular, when the squared residuals from otherwise well-specified linear models were regressed on their own lagged squared values, the regression coefficients often turned out to be highly significant. Engle realized that this was not actually a test for bilinearity but rather the optimal LM test for some other nonlinear model. Putting this together, Engle brought forth the ARCH model.

The particular ARCH(p) model first analysed and estimated by Engle (1982) may be succinctly expressed as

$$y_t = m_t + \varepsilon_t, \quad \varepsilon_t | \mathrm{I}_{t-1} \sim \mathrm{N}(0, \ h_t), \quad \text{and} \quad h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_p \varepsilon_{t-p}^2,$$

where I_{t-1} refers to the set of information available at time $t-1$, m_t denotes the conditional mean of the y_t time series, and all of the $\alpha_0, \ldots, \alpha_p$ parameters are restricted to be non-negative. The first equation for the conditional mean is, of course, completely standard (in his original application to UK consumer prices Engle used an error correction model for the mean). However, the key difference – Engle's brilliant new insight – comes from recognizing that even though the residuals, ε_t, must be serially uncorrelated, their conditional variance, and therefore the conditional variance of y_t, need not be constant but may in fact be predictable. Moreover, by explicit parameterizing h_t as a function of the past squared residuals and by assuming conditional normality, the joint density for all of the observations, say y_t, $t = 1, 2, \ldots, T$, may easily be evaluated through a prediction error decomposition type argument, and the log likelihood function maximized with respect to all of the model parameters, in turn resulting in a time series of positively serially correlated conditional variance estimates, $\hat{h}_t$, $t = 1, 2, \ldots, T$ (that is, estimates of inflationary uncertainty in Engle's original application).

While Engle's initial work and empirical applications of the ARCH model were rooted in macroeconomics, the model has shone most brightly in the area of finance. Since Mandelbrot's work in the early 1960s on the behaviour of speculative prices, it had been recognized that, even though most returns are approximately serially uncorrelated (at least over shorter daily or weekly horizons), 'large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes' (Mandelbrot 1963). However, the empirical finance literature up until the mid-1980s had largely ignored this fact, focusing instead on best characterizing the unconditional return distributions. Meanwhile, Engle soon realized that the ARCH model was ideally suited to this type of data: little, or no, serial correlation in the mean, but strong serial correlation in the second moments. Moreover, the ability to directly quantify the risk through a parametric model for the conditional variance, or more generally the conditional covariance matrix, for the returns strikes directly at the heart of the risk-return trade-off central to asset pricing finance. Consequently, Engle quickly shifted the focus of his research agenda to finance. Over the next 20 years, along with his many students and other collaborators, he developed numerous refinements to the basic ARCH model described above designed to account better for specific features of the data and/or questions of economic import: richer ARMA-type representations for the variance, including unit-root and long-memory type dependencies, models in which the variance directly influences the conditional mean, asymmetries or leverage effects in the variance, alternative parametric and non-parametric conditional distributions in place of the normal, multivariate factor models and cointegration in variance, to mention but a few. The corresponding long list of new

acronyms is also legendary: ARCH-M, GARCH, IGARCH, EGARCH, TARCH, GJR-GARCH, NARCH, QARCH, STARCH, VGARCH, SWARCH, FIGARCH – the list goes on. Empirical applications of these models have in turn resulted in many important new insights into the pricing and hedging of financial instruments and functioning of financial markets, and it is no exaggeration to say that the day-to-day risk management and monitoring in financial institutions have been completely altered by the advent of the ARCH class of models.

Not one to rest on his laurels, Engle continues to push forward the research frontier in financial econometrics. Most recently he has worked extensively on new methods for analysing ultra high-frequency, or tick-by-tick, financial data. In particular, whereas most procedures in time series econometrics, including most of Engle's own earlier work, are explicitly designed for modelling discretely sampled equidistant observations, high-frequency financial data are typically not observed at fixed time intervals. Engle's recent Autoregressive Conditional Duration (ACD) model, which derives many of its statistical properties from the ARCH class of models, provides a particularly convenient way of accommodating this feature by explicitly modelling the times between observations as a serially correlated process. His Dynamic Conditional Correlation (DCC) model, which allows for the estimation of large-scale dynamic covariance matrices, represents another recent noteworthy advance. In keeping with his trademark, this latest research represents the perfect blend between sophisticated yet simple-to-implement econometric techniques explicitly designed for answering genuinely interesting economic questions. Like most of his research since the 1970s, his latest work has already found widespread use both inside and outside academia, and spurred a number of ongoing new developments by other researchers in the field.

In addition to the much-deserved recognition bestowed on him by the Nobel Prize Committee, Engle is a long-standing fellow of the Econometric Society, of the American Statistical Association, and of the American Academy of Arts and Sciences. He is also an excellent speaker, and he has a long list of invited talks and keynote addresses to his name, including the prestigious A. W. Philips and Fisher-Schultz lectures sponsored by the Econometric Society. (For a more in-depth discussion of Engle's work along with some personal reflections, see Diebold 2003, 2004.)

In conclusion, it is simply impossible to imagine what the field of time series econometrics, let alone the new field of financial econometrics, would have looked like today had it not been for Engle's seminal contributions, both direct and indirect, through the substantial subsequent research programmes his work has helped stimulate. But Engle isn't merely one of the greatest econometricians of his time. He has a wide range of other interests and talents. For example, he is an outstanding ice skater, having competed at the US national level, finishing second in the 1996 and 1999 ice dancing championship competition.

See Also

- ▶ ARCH Models
- ▶ Cointegration
- ▶ Econometrics
- ▶ Extremal Quantiles and Value-at-Risk
- ▶ Forecasting
- ▶ Granger, Clive W. J. (1934–2009)
- ▶ Measurement Error Models
- ▶ Risk

Selected Works

1972. (With F. Fisher, J. Harris and J. Rothenberg.) An econometric simulation model of intra-metropolitan housing location: housing, business, transportation and local government. *American Economic Review* 62: 87–97.

1973. Band spectrum regression. *International Economic Review* 15: 1–11.

1976. Interpreting spectral analysis in terms of time domain models. *Annals of Economic and Social Measurement* 5: 89–109.

1979. (With C. Granger, R. Ramanathan and A. Andersen.) Residential load curves and time-of-day pricing: An econometric analysis. *Journal of Econometrics* 9: 13–32.

1981. (With M. Watson.) A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association* 76: 774–781.

1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* 50: 987–1008.

1983a. (With D. Hendry and J. Richard.) Exogeneity. *Econometrica* 51: 277–304.

1983b. (With M. Watson). Alternative algorithms for the estimation of dynamic factor, MIMIC, and varying coefficient regression models. *Journal of Econometrics* 23: 385–400.

1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of econometrics*, Vol. 3, ed. Z. Griliches and M. Intrilligator. Amsterdam: North-Holland.

1986a. (With C. Granger, J. Rice and A. Weiss.) Semi-parametric estimation of the relation between weather and electricity demand. *Journal of the American Statistical Association* 81: 310–320.

1986b. (With T. Bollerslev.) Modeling the persistence in conditional variances. *Econometric Reviews* 5: 1–50.

1987a. (With C. Granger.) Co-integration and error correction: Representation estimation and testing. *Econometrica* 55: 251–276.

1987b. (With D. Lilien and R. Robins.) Estimation of time varying risk premia in the term structure: The ARCH-M model. *Econometrica* 55: 391–407.

1987c. (With S. Yoo.) Forecasting and testing in co-integrated systems. *Journal of Econometrics* 35: 143–159.

1988. (With T. Bollerslev and J. Wooldridge.) A capital asset pricing model with time varying covariances. *Journal of Political Economy* 96: 116–131.

1990a. (With T. Ito and W. Lin.) Meteor showers or heat waves? Heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica* 58: 525–542.

1990b. Asset pricing with a factor ARCH covariance structure: Empirical estimates for Treasury bills (with V. Ng and M. Rothschild). *Journal of Econometrics* 45: 213–237.

1990c. Seasonal integration and cointegration (with S. Hylleberg, C.W.J. Granger and B.S. Yoo). *Journal of Econometrics* 40: 45–62.

1991a. (With C. Granger.) *Long run economic relations: Readings in cointegration*. Oxford: Oxford University Press.

1991b. (With G. Gonzales.) Semi-parametric ARCH models. *Journal of Business and Economic Statistics* 9: 345–359.

1992. (With C. Mustafa.) Implied ARCH models from options prices. *Journal of Econometrics* 52: 289–311.

1993a. (With T. Bollerslev.) Common persistence in conditional variances. *Econometrica* 61: 167–186.

1993b. (With S. Kozicki). Testing for common features. *Journal of Business and Economic Statistics* 11: 369–380.

1994a. *Handbook of econometrics*, Vol. 4, ed. with D. McFadden. Amsterdam: North-Holland.

1994b. (With T. Bollerslev and D. Nelson.) ARCH models. In *Handbook of econometrics*, Vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

1995a. *ARCH: Selected readings*. Oxford: Oxford University Press.

1995b. (With K. Kroner.) Multivariate simultaneous ARCH. *Econometric Theory* 11: 122–150.

1998. (With J. Russell). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* 66, 1127–62.

1999. (With G. Lee). A permanent and transitory component model of stock return volatility. In *Cointegration, causality, and forecasting: A festschrift in honor of Clive W. J. Granger*, ed. R. Engle and H. White, 475–497. Oxford: Oxford University Press.

2000. The econometrics of ultra high frequency data. *Econometrica* 68: 1–22.

2002a. Dynamic conditional correlation: A simple class of multivariate GARCH models. *Journal*

of Business and Economic Statistics 20: 339–350.
2002b. (With J. Rosenberg.) Empirical pricing kernels. *Journal of Financial Economics* 64: 341–372.
2004. Risk and volatility: Econometric models and financial practice. *American Economic Review* 94: 405–420.

Bibliography

Diebold, F. 2003. The ET interview: Professor Robert F. Engle. *Econometric Theory* 19: 1159–1193.
Diebold, F. 2004. The Nobel memorial for Robert F. Engle. *Scandinavian Journal of Economics* 106: 165–185.
Mandelbrot, B. 1963. The variation in certain speculative prices. *Journal of Business* 36: 394–419.

English School of Political Economy

A. M. C. Waterman

Abstract

The 'English School' of political economy describes the tradition of economic thought that began with Malthus, and included Henry Thornton, Chalmers, James Mill, Torrens, West, Ricardo, and Thomas Tooke in the first generation; Whately, Senior, McCulloch and J.S. Mill in the second; and the Fawcetts, Cairnes, Jevons, Bagehot, Foxwell, Sidgwick, J.N. Keynes and Nicholson in the third. J.-B. Say was an honorary member. Karl Marx identified his own work with that school. Its most important production was J.S. Mill's *Principles of Political Economy,* which continued to be used as a textbook until the mid-20th century.

Keywords

American economic association; Anderson, J.; Bagehot, W.; Bentham, J.; Böhm- Bawerk, E. von; Boisguilbert, P. de; Brougham, H.; Bullion committee; Cairnes, J. E.; Cassel, G.; Chadwick, E.; Chalmers, T.; Christian political economy; Clark, J. B.; Classical economics; Cliffe Leslie, T. E.; Comparative advantage; Condillac, E. B. de; Diminishing returns; Edgeworth, F. Y.; English school of political economy; Factor substitution; Fawcett, H.; Fawcett, M. G.; Fisher, I.; Foxwell, H. S.; Free trade; Harrod, R. F.; Horner, F.; Hume, D.; Increasing returns to scale; Jeffrey, F.; Jevons, W. S.; Jones, R.; Keynes, J. N.; Labour theory of value; Lauderdale, eighth earl of; Lloyd, W. F.; Longfield, M.; Malthus, T. R.; Mandeville, B.; Marginal revolution; Marginal utility; Marshall, A.; Marx, K. H.; Mathematics and economics; Mayerne- Turquet, L. de; McCulloch, J. R.; Menger, C.; Mercantilism; Mill, J.; Mill, J. S.; Montchrétien, A. de; Natural price; Neutrality of money; Nicholson, J. S.; Norman, G. W.; Pareto, V.; Pigou, A. C.; Political economy; Political economy club; Poor law; Productive and unproductive labour; Quesnay, F.; Ricardo, D.; Robbins, L. C.; Say, J.-B.; Say's identity; Say's law; Scrope, G. P.; Senior, N. W.; Sidgwick, H.; Smith, A.; Smith, S.; Sraffa, P.; Stationary state; Steuart, J.; Stewart, D.; Sumner, J. B.; Surplus; Thornton, H.; Thornton, W. T.; Thünen, J. H. von; Tooke, T.; Torrens, R.; Turgot, A. R. J.; Utilitarianism; Walker, F. A.; Walras, L.; Wayland, F.; Weiser; West, E.; Whately, R.; Whewell, W.; Wicksell, J. G. K.; Wicksteed, P. H.; Young, A. A.

JEL Classifications

B1

The 'English School' of political economy comprises all major British economists of the 19th century, together with J.-B. Say and perhaps Karl Marx.

'Important changes have taken place in the meaning of the term "political economy," as used by leading writers, since it was first employed', wrote Henry Sidgwick in Palgrave's original *Dictionary of Political Economy* (Palgrave 1899, pp. 128–9). As first used by Mayerne-Turquet and Montchrétien, *'œconomie*

politique' signified an attempt to extend the art of estate management to the entire kingdom of Louis XIII and his successors (Waterman 2004, p. 225). This usage, generalized to mean a 'system' of policy designed to 'increase the riches and power' of a country (Smith 1776, I.xi.n.1; II.5.31; IV.1.3) remained current until the end of the 18th century and was so employed by Steuart.

Adam Smith disliked the usage because of its implicit mercantilism. He recognized it, but proposed a better definition. 'What is properly called Political Œconomy' is 'a branch of the science of a statesman or legislator': namely *'an inquiry',* which is in principle disinterested and open-ended, into *'the nature and causes of the wealth of nations'* (Smith 1776, IV. intro; IV.ix.38; emphasis added). The prestige that *Wealth of Nations* quickly acquired, amplified by Dugald Stewart's widely influential Edinburgh lectures in the new science, redefined 'political economy' as a 'part of the science of human society' (Palgrave 1899, p. 129; cf. Winch 1983, who appears to disagree with this interpretation) and created a circle of younger thinkers committed both to criticizing and refining Smith's ideas and to propagating them among the governing classes. Though the *Edinburgh Review*, founded in 1802, was at first the principal means of propagation, most of the prime movers soon migrated to London, which from the second or third decade of the 19th century became the home of what was soon called the 'English School'.

It is important to recognize that to describe the small community of anglophone political economists in the 1820s and after as a 'school' is to imply neither a quasi- apostolic succession of doctrine in some leading university nor a closed shop of experts defined by their adherence to any orthodoxy. It is rather the fact, as T.S. Eliot observed of all such intellectual circles in general, that 'they are driven to each other's company by their common dissimilarity from everybody else, and by the fact that they find each other the most profitable people to disagree with' (Kojecky 1971, p. 244). Members of the English School, like all subsequent economists, were notorious for their disagreements, both with Adam Smith and with each other. But they did not find it profitable to disagree with hostile critics of their enterprise, such as the Lake Poets (from whom they were all indeed markedly 'dissimilar'), because the latter chose not to acquire the viewpoint and vocabulary of the new, political-economy conversation, but resorted rather to the idioms of a very different conversation: that of Romantic aesthetics and non-utilitarian ethics.

In attending to the conversation of the English School it is necessary first to establish its identity, secondly to consider its members and its literature, and thirdly to distinguish its chief analytical features, especially as these differ both from the economic thought that preceded it and from economics of the present day. Finally, since the boundary between the political economy of the English School and what is generally thought of as 'modern' economics is vague and permeable, some attention should be paid to continuity and 'revolution', if any.

Identity of the English School of Political Economy

Writing of 'the English School of Political Economy' in the original Palgrave dictionary, James Bonar (Palgrave 1894, p. 730) observed that 'The English writers on political economy before Adam Smith do not at any time present the marks of a "school" properly so called.' What Bonar called 'Modern Economics' – meaning 'political economy' in the new, Smithian sense – he then divided into four periods headed respectively: Adam Smith, Malthus and Ricardo, John Stuart Mill, and W.S. Jevons; with all other authors subsumed under these canonical names.

Adam Smith was not an Englishman and he died before Malthus and Ricardo had begun to write. Though the *Edinburgh Review* (Anon 1837, p. 73) referred to the English School as 'the school of which Adam Smith was the founder', this is Caledonian hyperbole. Smith founded no 'school'. His most influential disciple, Dugald Stewart, was the intermediary between Smith and those the *Edinburgh Review* more accurately described later in the article as the 'followers of Dr Smith' practising 'Political Economy, using

the word in the sense of Ricardo and Malthus' (Anon 1837, pp. 77, 79). Subject to this important qualification, Bonar's chronology is helpful. Roughly speaking the English School lasted for about three generations. The first generation, from 1798 to the 1830s is that of which Malthus, Henry Thornton, Chalmers, James Mill, Torrens, West, Ricardo, and Thomas Tooke are now the best remembered. A second generation, whose members were active in some cases before 1830, but who flourished for the most part until the 1860s or even later, included Whately, Senior, McCulloch and J.S. Mill. Political economy of the English School never really died out. It changed, very gradually and almost imperceptibly, into the international, professionalized 'economics' of the mid-20th century. Yet a third and last generation can be detected – and was in fact detected in the 1890s – which included W.T. Thornton, the Fawcetts, Cairnes, Jevons, Bagehot, Foxwell, Sidgwick, J.N. Keynes and Nicholson. The positions of Marshall, Edgeworth and Wicksteed are problematic and will be considered below.

The English School was recognized by its difference from 'the foreign school' (Anon 1837, p. 77) which included Sismondi, Cherbuliez and Villeneuve, but not J.-B. Say who from the first was deemed an honorary Englishman. The English writers distinguished 'the art of government' from the 'science' of political economy. With respect to the former, the latter is 'only one of many subservient sciences; which involves the consideration only of motives, of which the desire for wealth is only one among many, and aims at objects to which the possession of wealth is only a subordinate means' (Senior 1836, pp. 129–30). The foreign writers rejected this minimalist construal, labelled it 'chrematistics' or 'chrysology', and continued to maintain that political economy embraces both the art and the science of government.

The incipient distinction between 'art' and 'science' seemed to imply that any practitioner of the latter must abstain – qua political economist – from political judgements. His analytical conclusions, being strictly positive and abstracted from ethical considerations, 'do not authorize him in adding a single syllable of advice'. His business is 'neither to recommend nor to dissuade, but to state general principles which it is fatal to neglect' (Senior 1836). McCulloch for one strongly disagreed with Senior on this point: the general principles, he thought, had already been completely enounced by Ricardo. What remained was 'to exhibit some of their more important applications' (McCulloch 1843, p. vi). Though Senior's view of the scope of political economy was tidied up and assimilated by the end of the 19th century (J.N. Keynes 1891), all members of the 'school' were agreed at the outset on at least one most important 'application'. The 'great principles of free exchange and natural distribution' that Smith had developed from 'the philosophers of the Continent' (that is, Quesnay, Turgot and so on) showed it to be economically unprofitable 'for the legislature to intermeddle' with trade and income distribution (Anon 1837, pp. 80, 78). Though Cairnes later averred that 'political economy has nothing to do with laisser faire', Sidgwick thought this 'too daring a paradox'.

> There can be no doubt that the interest of Adam Smith's book for ordinary readers is largely due to the decisiveness with which he offers to statesmen the kind of practical counsels which, according to Senior and Cairnes, he ought carefully to have abstained from giving. (Palgrave 1899, pp. 130–1)

Rightly or wrongly, the political economy of the English School was associated in the popular mind with free trade and attacks on corporate privilege, and was denounced for these disturbing ideas by a wide variety of hostile critics.

Both the methodological tendencies of the new science and its 'more important applications' owed much to Dugald Stewart: the former to his influential *Philosophy of the Human Mind* (1792, 1814, 1827) the latter to his annual public lectures at the University of Edinburgh, beginning in the winter of 1800/1.

Though preferring a broader definition of 'political economy' than that of either Smith or his English followers, and emphasizing the historical character of economic knowledge, Stewart argued in *Human Mind* that the hypothetical and a priori reasoning so characteristic of what he called the 'new science' – and which became

one of the hallmarks of the English School – was perfectly legitimate, and compatible with the testing of theories against experience (Fontana 1985, pp. 99–102; Waterman 2004, ch. 8).

Stewart's Edinburgh lectures were crucial in what a recent author has aptly called 'the process of Anglicisation of Scottish thought after 1790' (Fontana 1985, p. 9). Not only were they attended by Jeffrey, Horner, Brougham, Chalmers and the newly arrived Englishman, Sydney Smith, all of whom were influential in propagating political economy; Pryme (1823, p. vii) records that they 'attracted so much attention that several members of our own university [namely, Cambridge] went from the South of England to pass the Winter at Edinburgh, for the purpose of attending them': one of these seems to have been John Bird Sumner (Waterman 1991a, pp. 159–60). According to a later account, 'a wave of young Englishmen . . . went North in lieu of the grand tour made impossible by the renewal of war' (Checkland 1951, p. 43). Though the lectures were diffuse and circumspect, their underlying message was that contained in an early paper that Adam Smith had entrusted to Stewart before his death:

> Little else is required to carry a state to the highest degree of opulence from the lowest barbarism, but peace, easy taxes and a tolerable administration of justice; all the rest being brought about by the natural course of things (Smith 1755). (Winch 1996, p. 90)

Leading members of Stewart's circle – Jeffrey, Horner, Brougham, and Sydney Smith – founded the *Edinburgh Review* to urge this message upon the Holland House Whigs from whom they hoped to receive patronage. First Smith, then Jeffrey, served as editor until 1829, when replaced by McVey Napier. By that date its contributors on political economy had included all the leading members of the English School save Ricardo (who declined out of modesty, and who died in 1823): Malthus, James Mill, Chalmers, Torrens and McCulloch (Fontana 1985, p. 8).

Of these authors, all save Chalmers were members of the Political Economy Club, a London dining club founded in 1821 which, in addition to Malthus, Mill and Torrens, included from the outset Ricardo, George Warde Norman and Thomas Tooke. J.-B. Say was elected as an Honorary Foreign Member in 1822, the only such member until 1919. McCulloch was elected in 1829, shortly after his migration from Scotland; Senior (in 1823), Pryme (1828) and Whately (1831) were elected as Honorary Members by virtue of their professorships in political economy. Cairnes (1862), Cliffe Leslie (1862), Fawcett (1862), Jevons (1873), Foxwell (1882), Marshall (1886), Nicholson (1888) and Edgeworth (1891) were all subsequently elected under this rule. Among those political economists now remembered as influential authors of the English School, only Henry Thornton, Sir Edward West, Archbishop J. B. Sumner, Thomas Chalmers, Poulett Scrope, and Richard Jones were never members of the Club: Thornton because he died in 1815, West because he went to India, Chalmers because he stayed in Scotland, and Sumner because he announced in 1818 – to Ricardo's regret – that he intended to give up political economy for the study of theology (Waterman 1991a, p. 157). Scrope and Jones were on the outer edge of the 'School'.

It has been suggested that the English School was a 'scientific community' of which the Political Economy Club was a 'vital hub' (O'Brien 2004, pp. 12–13). There is merit in this suggestion, but it should be recognized that the original purpose of the club, though including the 'mutual instruction' of members, was chiefly propagandist: 'the diffusion amongst others of the just principles of Political Economy' and

> to watch carefully the proceedings of the Press, and to ascertain if any doctrines hostile to sound views on Political Economy have been propagated . . . to refute such erroneous doctrines, and counteract their influence . . . and to limit the influence of hurtful publications. (Political Economy Club 1921, p. 375)

Many members were Whig or liberal statesmen who knew a 'hurtful publication' when they saw one: 52 of the 115 elected between 1821 and 1870 sat in either the upper or lower House of Parliament; and included Lord Althorp, the Marquis of Landsdown (a descendent of Sir William Petty), Earl Grey and W. E. Gladstone. Fetter (1980) has

documented the activities of 'the economists in Parliament'.

Almost from the first there was a desire by the Club to recognize and foster the academic study of political economy. Though there had been high-level economic analysis at British universities before the end of the 18th century, it was but a small ingredient of 'moral and political philosophy' (for example, see Waterman 1995) and never known as 'political economy'. But in the decade of the 1820s chairs in political economy were established in Oxford, London and Cambridge and their incumbents immediately co-opted (Checkland 1951).

We may therefore identify the English School roughly speaking as that subset of Political Economy Club members in the 19th century who published and disputed with each other on the subject, together with half a dozen or so other major authors who at some time or other were part of their conversation. Despite its name, several leading members were Scotch immigrants, and it included one Frenchman. Though Karl Marx lived in London from 1848 and thoroughly digested the literature of anglophone political economy over the next two decades, he was not known or recognized by the Club. But in the 'Afterword' to the second German edition of *Capital* (Marx 1873, vol. 1, p. 26) he explicitly identified his own work, in method at least, with that of the English School.

Literature of the English School

Literature of the English School begins with Malthus's first *Essay on Population* (1798). For as an unintended consequence of his Whiggish polemic against Godwin's (1793) romantic anarchism, Malthus analysed the effect of population growth under land scarcity to show what was later called 'diminishing returns' (Stigler 1952). Though diminishing returns in agriculture had been identified by Steuart (1767) and Turgot (1768), and had actually been used by Anderson (1777) to adumbrate the 'Ricardian' theory of rent, the concept was not integrated into 18th-century economic thought. Notwithstanding Samuelson's influential interpretation, land scarcity plays little or no analytical part in *Wealth of Nations* (Samuelson 1978; cf. Hollander 1998; Waterman 1999). When Malthus (1815a), West (1815), Torrens (1815) and Ricardo (1815) worked out the implications of Malthus (1798) they believed that they were correcting Smith and saying something new and important (McCulloch 1845, p. 68). Diminishing returns immediately became part of the hard core of the so-called classical political economy of the English School.

Ricardo made diminishing returns in agriculture the cornerstone of his *Principles* (1817), combined it with 'Malthusian' population theory, Smith's account of accumulation and growth, and an ad hoc '93 % Labor Theory of Value' (Stigler 1958) to produce a complete account of value, distribution and growth in a two- sector market economy. The labour theory of value (LTV) was also the key concept in Ricardo's rigorous and elegant analysis of comparative advantage in international trade. Looking back 30 years later, McCulloch (1845, p. 16) called the LTV 'the fundamental theorem of the science of value'. An authoritative and exhaustive account of Ricardo's contribution – which it treats, à la McCulloch, as virtually identical with 'classical economics' – appeared in the first edition of *The New Palgrave Dictionary of Economics* (Blaug 1987).

In addition to the above works, the 'English' literature that already existed by the time the Political Economy Club was founded in 1821 included Malthus's (1800) *High Price of Provisions,* which formally specified a demand function of price and inaugurated the supply-and-demand value theory that eventually 'won out' over the Ricardo–Marx LTV (Smith 1956; Schumpeter 1954, p. 48) which it generalizes, Thornton's (1802) *Paper Currency,* which analysed the macroeconomic relations between monetary and real variables in a manner reinvented by Wicksell a century later, and numerous pamphlets by many authors on monetary questions provoked by the Parliamentary Bullion Committee of 1810. It was this controversy that brought Malthus and Ricardo together, and which seems to have been a catalyst for the nascent

'scientific community'. The pre-1821 literature also includes J.-B. Say's (1803) *Traité d'économie politique*; Lauderdale and Maitland (1804) *Inquiry,* dismissed by McCulloch (1845, p. 15) as without value; Chalmers's (1808) strikingly original but completely neglected *Nature and Stability of National Resources* (see Waterman 1991b); Malthus's (1815b) heretical pamphlet, 'Restricting the Importation of Foreign Corn', which led to his excommunication by the *Edinburgh Review* (Fontana 1985, p. 75); J. B. Sumner's (1816) *Records of the Creation* that Ricardo (1951–73, vol. 7, pp. 247–8) deemed a 'clever book' and which McCulloch (1845, p. 261 described as 'an excellent work'; the fifth edition of Malthus's Essay on Population (1817) substantially modified as a result of Sumner's arguments; Mrs Marcet's (1817) influential work of popularization, *Conversations on Political Economy;* and Copleston's (1819a, b) two brilliant and penetrating *Letters to Peel* that grasped more clearly than Malthus himself the connection between population and poverty, and between the latter and inflation of the currency – and which Ricardo so admired that he made a detailed paragraph-by-paragraph summary (Waterman 1991a, pp. 186–95; Hollander 1932, p. 135–45). Finally, shortly before or just after the first meeting of the Club there appeared important monographs by three of the founding fathers: Malthus's (1820) *Principles,* which quarrelled with Ricardo over value theory and put forward a heterodox macroeconomics of 'general gluts' that Keynes was later to find so appealing, James Mill's (1821) *Elements of Political Economy,* and Torrens's (1821) long undervalued *Essay on the Production of Wealth.*

It is apparent that during the first two decades of the 19th century, and for a further ten years or more, Malthus was at the centre of the political-economy conversation of the English School. This fact has been obscured by the excessive attention paid to Ricardo by those eager to praise or blame him for present-day economics, and by textbook authors wanting a handle on which to hang a student- friendly chapter on 'classical economics'. A long process of reappraisal, beginning with J. M. Keynes's (1972, vol. 10) biographical essay of 1933, has gradually restored the true picture (Waterman 1998). Donald Winch's (1996) *Riches and Poverty* is the latest and most authoritative intellectual history of political economy, covering the period 1750–1834. Nearly half his book is concerned with Malthus. Ricardo, 'treated largely as a foil to Malthus' (Winch 1996, p. 15) gets a few scattered references. Samuel Hollander's (1997) magisterial *Economics of Thomas Robert Malthus* shows that the analytical differences between Malthus and Ricardo have been exaggerated, and that the former was a theoretician of the same order, and of at least as much historical importance as the latter.

Malthus was central because the first *Essay* began a century-long transformation of 'political economy' (the science of wealth) into 'economics' (the science of scarcity). The theological implications of this, totally ignored by most historians, are a vital part of the intellectual context of the English School. Economic thought of the 18th century was believed by all to be wholly compatible with Christianity. But the seeming inevitability of 'misery' or 'vice' produced by human fecundity and resource scarcity challenges the goodness of God; and the political economy of Malthus and Ricardo was therefore condemned as 'hostile to religion'. For most of the 19th century, England was both officially and actually a Christian society. In such a society it is part of the duty of a scientist – essential if his work is to receive serious attention – to reconcile his findings with Christian theology. Malthus attempted this in 1798 and failed. His failure stimulated an important branch of the literature of the English School now known as 'Christian Political Economy' (Waterman 1991a). Works by William Paley (1802), by Malthus himself (1803, 1817), and by J. B. Sumner (1816) who eventually became Archbishop of Canterbury, demonstrated that the new science could be co-opted as theodicy; and even better, be used to demonstrate the benevolent 'design' of the Creator. The approval that Ricardo and McCulloch evinced for Sumner's 'clever book' had less to do with their own religious convictions than with their relief that political economy had been convincingly defended against the damaging charge of irreligion.

Quite different circumstances in the 1820s revived the need to defend political economy against religion, and created a new need: to defend religion against political economy. Jeremy Bentham, James Mill and other Benthamites, who were later called the 'Philosophic Radicals', founded the *Westminster Review* in 1824 to propagate a 'radical' reformism as against the Whiggish reformism of the *Edinburgh.* Anticlerical and at times anti-religious, the radicals hijacked political economy to mount a strictly utilitarian attack on the Establishment in Church and State. Animated by James Mill's puritanical hatred of the Arts, the *Westminster* compounded the injury by gratuitous attacks on the Lake Poets and other romantic authors. Influential Tories at the two universities (then exclusively Anglican) were alarmed, and opposition was made to the teaching of, and the establishment of chairs in, political economy. In this crisis, both political economy and Christian theology were authenticated and insulated against mutual encroachment by two Oxford men, Richard Whately, a former pupil and friend of Copleston, and Nassau Senior, Whately's former pupil and friend (Waterman 1991a, pp. 196–215).

Whately engineered the election of Senior as first Drummond Professor of Political Economy in 1826, and accepted the chair himself when it fell vacant in 1830. His seminal Introductory Lectures (1831) argued for an epistemological demarcation between 'religious and 'scientific' knowledge; and explained how, like all scientific knowledge, political economy depends upon both a priori deduction and the possibility of falsification. Whately thus established the methodological tradition of the English School that runs through Senior, J.S. Mill, J.N. Keynes and Lionel Robbins. Pietro Corsi (1987) has shown that Whately's philosophical apparatus was based on Dugald Stewart's *Philosophy of the Human Mind,* transmitted to Oxford through the friendship between Stewart and Copleston created by the migration from Edinburgh to Oxford in 1799 of J.W. Ward, 1st Earl of Dudley.

Whately's decisive intervention healed a potentially disastrous schism in the young 'scientific community' between Benthamite radicals and Malthusian Whigs. Elections to the Political Economy Club in the 1820s and 1830s included both Whigs and radicals and even the liberal Tory, Lord Althorp. When McVey Napier edited the 1824 *Supplement to the Encyclopaedia Britannica* he commissioned articles on political economy from Malthus and Sumner on the one hand, and from Mill and McCulloch on the other. (Ricardo's contribution, on the Funding System, was posthumous.) The Royal Commission on the Poor Laws (1832) which included Sumner, then Bishop of Chester, united all in the common cause once again. Malthus was the most important witness. The report, which led to the Poor Law Amendment Act (1834), was jointly written by the Benthamite Chadwick and the Whatelian Senior, and was based on Copleston's (1819b, p. 28) crucial distinction between 'propagation' and 'preservation' of human life.

One of the most interesting, certainly the most revealing, contributions to literature of the English School is McCulloch's compendious *Literature of Political Economy* (1845) which appeared about halfway through the life of the 'school'. The usual English and Scotch authors from Mun and Petty are listed, and many of their works praised or censured in light of McCulloch's doctrinal preconceptions. Malthus is predictably belittled. All the leading French authors of the 18th and early 19th centuries appear save Boisguilbert and Cournot. Condillac's pathbreaking *Le Commerce et le gouvernement* (1776) is dismissed with a patronizing comment of J.-B. Say (McCulloch 1845, p. 63; cf. Eltis and Eltis 1997, pp. 30–4). Considerable respect is paid to Italian authors (McCulloch 1845, pp. 28–31, 86), but the Spanish are written off as intellectually impotent until Napoleon's invasion (1845, pp. 31–2, 326). McCulloch seems never to have heard of Thünen, and no other German author is mentioned. Omissions of anglophone authors are equally telling. Whewell's pioneering mathematical economics is ignored, presumably for the same reason as the omission of Cournot and Thünen. Dugald Stewart is cited merely as a biographer of Adam Smith and Robertson (1845, pp. 8, 104, 162). McCulloch seems not have read or understood either Chalmers (1808) or Copleston

(1818), nor to have grasped the analytical significance of Malthus (1800). Everything is viewed through the powerful but slightly distorting lenses of Adam Smith and Ricardo.

Three years later there appeared the single most important production of the School: J.S. Mill's *Principles of Political Economy* (1848), perceptively reviewed by Bagehot (1848) among many others. Mill's *Principles* is the definitive statement of the English School of political economy. It went through seven editions in the author's lifetime; the 1909 scholarly edition by Ashley was based on the seventh (1871), and may be taken as the terminus ad quem of the English School. For though Mill continued to be the principal textbook in political economy until the 1930s at many universities throughout the English-speaking world, Anglophone economic literature of the 20th century gradually became less insular (Palgrave 1894, p. 735) and was formed in the cautiously new idiom of Marshall and Pigou, with at least some peripheral awareness of Jevons and Edgeworth, Walras and Pareto, Weiser and Böhm-Bawerk, Cassel and Wicksell, J.B. Clark and Fisher.

Though Mill dominated, there were many other significant contributions to the literature in the last third of the 19th century. Henry Fawcett's *Manual of Political Economy* (1863) encapsulated Mill's *Principles* for faint-hearted undergraduates; his wife's even more elementary *Political Economy for Beginners* (1870) went through ten editions over the next 41 years. W.T. Thornton's *On Labour* (1869) introduced the concept of multiple equilibria, as Mill (1869, p. 637) admitted. Cairnes's *Leading Principles* first appeared in 1874, Cliffe Leslie's *Essays* in 1879, Bagehot's posthumous *Economic Studies* in 1880, and Henry Sidgwick's *Political Economy* in 1883. Sidgwick's importance in the incipient 'Cambridge' mutation of the English School has lately been documented (Backhouse 2006). J.N Keynes's classic *Scope and Method* first appeared in 1891. Perhaps the last major production of the English School was J. Shield Nicholson's three-volume *Principles of Political Economy* (1893–1901), a basically Millian exposition with the occasional bow to Marshall, used as a textbook in many parts of the British Empire in the early 20th century. Nicholson's appears to be the last widely read work of political economy to consider explicitly the relation between that science and Christian theology (1893–1901, vol. 3, ch. 20).

Stanley Jevons (1871, p. 275) went out of his way to challenge 'the noxious influence of authority' in the English School, above that of Mill. Though elected to the Political Economy Club as a professor in 1873 and as an Ordinary Member in 1882 (the year of his death), he was therefore handled with caution by his fellow-economists – including the powerfully influential Marshall. Whilst crediting him with the intellectual defeat of Ricardian and Marxian value theory, Bonar (Palgrave 1894, p. 735) thought that 'the ideas of Jevons have had greater power since his death than during his life'. Jevons and his two most creative English followers, Edgeworth and Wicksteed, were 'often spoken of as a school by itself, the mathematical school' (Palgrave 1894). The original Palgrave article on 'Recent Developments of Political Economy' (Palgrave 1894, p. 148) alludes to Jevons's *State in Relation to Labour* (1882) but ignores his *Theory of Political Economy* (1871).

Literature of the English School was augmented and popularized by *The Economist* newspaper, founded in 1843 and edited by Walter Bagehot from 1860 to 1877, which, like the Political Economy Club, sought to relate economic analysis to public policy in the spirit of Adam Smith. That literature may be said to have culminated in the three-volume *Dictionary of Political Economy* (1894–1899) edited by R.H. Inglis Palgrave.

Some Analytical Features of the English School

Political economists of the English School inherited much of their economic analysis from their 18th-century predecessors, especially Cantillon, Hume, Quesnay, Smith and Turgot. However, some features of their analysis were as 'novel' as any idea ever is in the social sciences.

And despite loose talk about a 'marginal revolution', much of their analysis, both what they inherited and what they originated, has become part of the stock-in-trade of present-day economics. The standard account by D.P. O'Brien (2004) should be supplemented by S.J. Peart's and D. Levy's (2003) review of the period 1830–1870, which considers catallactics, methodological egalitarianism and the new ideological alliance – a mutation of the old Whig-Liberal orthodoxy – between political economists and reformist Evangelicals in the Church of England.

The central conception of 18th-century economic thought was that of a surplus of production in one period over and above what is necessary (as inputs into production) to sustain that level of production in the next. The agricultural sector is an obvious source of the surplus since land normally produces more than the (food) cost of necessary labour and capital inputs. But Smith generalized the concept to include all produced goods capable of use as inputs. Masters incur production costs in advance, hence control the entire output at the end of the process. Some of this they consume either directly, or in the employment of unproductive labour. The remainder is used to feed and equip productive labour. This unconsumed portion of output is the (circulating) capital stock of a master, firm or community, the growth, stationarity or decay of which depends on a psychological propensity of masters: the extent of their 'frugality' or parsimony (Eltis 2000, pp. 75–100). These ideas, and the necessarily dynamic analytical framework they imply, were taken for granted by most the English School despite its seeming incompatibility with such other conceptions as comparative advantage in trade (Blaug 1987, vol. 1, pp. 439–42). Other characteristically 18th-century ideas accepted by 'the followers of Dr Smith' included that of a labour supply perfectly elastic in the (Malthusian) long period at a socially determined zero-population-growth real wage; enough factor mobility to produce uniform rates of wages and profit throughout the economy; a negative relation between the real wage and the rate of profit; a positive relation between the general price level and the stock of money, and the Cantillon–Hume price-specie- flow mechanism of international monetary adjustment which follows from that relation. Most accepted Smith's account of natural prices that correspond, more or less, to Marshall's long-period equilibrium prices, but O'Brien (2004, ch. 4) has shown in detail how much variation there was in this matter. Perhaps the most important 18th-century idea, certainly that which gave the English School its ideological momentum, was Boisguilbert's vision – derived from the Jansenist theology of Pierre Nicole and Jean Domat – of a self-regulating market economy driven by 'self-love' and producing some kind of social optimum at competitive equilibrium (Faccarello 1999). This powerful conception was transmitted by Mandeville, Cantillon and Quesnay and canonized by Smith in *Wealth of Nations*.

As we have seen, the English School made at least one sharp analytical break with 18th-century thought. The explicit incorporation of diminishing returns (though as yet in agricultural production only) created a fundamentally different view of the economic universe. Though all recognized increasing returns to scale (IRS) resulting from the division of labour, IRS plays a small or negligible part in the implicit growth models of Malthus and his successors (Eltis 2000). The salient feature of the new growth theory was rather a tendency for the rate of profit to fall: either because of rising costs in agriculture as in Malthus and Ricardo, or because of increasing capital intensity in manufactures as in Marx. In the former case, falling real factor payments retarded the growth of capital and labour, leading to a stationary state in the absence of technical progress. Samuelson (1978) has shown that the variable factor in agriculture was conceived as a single 'labor-cum-capital' unit, and though all 'classical' economists recognized the possibility of factor substitution especially in manufacturing, the capital–labour ratio was generally taken as a parameter. The same was true of technique. Improvements were seen to occur from time to time, and their effect upon wages, profits and employment analysed. Malthus, and perhaps some others, recognized that technical progress could become endogenous (Eltis 2000, pp. 150 ff.) and few if any of the English School

regarded it, as some do today, as 'manna from Heaven'. Two other new, or somewhat new, analytical features of the English School deserve note. The first is the LTV theory of comparative advantage, later improved by Mill's analysis of reciprocal demand. The second is Say's Law of Markets, which in its strong form (Say's identity) implies the neutrality of money (Blaug 1996, pp. 143–60). Whether Samuel Hollander (for example, 1987, pp. 6–7) is correct in maintaining that Ricardo and his contemporaries and successors, including Marx, recognized 'a fundamentally important core of general-equilibrium economics accounting for resource allocation in terms of the rationing function of relative prices' is still a matter of debate (Blaug 1987, vol. 1, pp. 442–3).

It is evident that most of these analytical characteristics, both those inherited from the 18th century and those that were new, have been transmitted to present-day economic thought. The obvious exception is the concept of a surplus with its concomitant distinction between 'productive' and 'unproductive' labour; though in the spirit of Feyerabend's (1988) methodological anarchism this venerable doctrine has lately been brought back to useful life (Bacon and Eltis 1976). For the most part however, present-day economists prefer to rely on a putatively constant-returns-to- scale (CRS) general equilibrium model that abstracts from time, and in which each factor-owner is paid the value of his factor's marginal product. The surplus is therefore regarded as a museum piece and left to heterodox Marxists and Sraffians (Walsh and Gram 1980; cf. Blaug 1987, vol. 1, pp. 440–2). It is important to recognize, however, that the eventual disappearance of the surplus in a neoclassical theory of distribution was brought about by an ever wider application of the marginal analysis originally applied by Steuart, Turgot, and Anderson, and then by Malthus, Ricardo and their contemporaries to agricultural production costs alone (Blaug 1987, vol. 1, p. 441). Authors of the next generation such as Longfield and Lloyd began the analysis of marginal utility (O'Brien 2004, pp. 119–22). Replacement of the dynamic surplus macroeconomics by a static general-equilibrium microeconomics dependent on universal CRS created perhaps the most significant analytical difference between political economists of the English School and the new professionalized economists of the early 20th century: an almost complete lack of interest among the latter in macroeconomics and growth theory. Not until Keynes's rediscovery of Malthus (Kates 1994) and Harrod's (1939) critique of Keynesian 'equilibrium' did these return to the theoretical agenda. As for Adam Smith's IRS, quietly forgotten by most of the English School – save Marshall – for most of the time and ignored by their successors, its reintroduction by Sraffa (1926) and Young (1928) has remained a thorn in the flesh for general equilibrium theorists.

Revolution and Continuity

Present-day 'economics' looks quite different from 'political economy' of the English School. Yet despite Samuelson's remarks about Marshall in *Foundations* (1947, pp. 6, 142, 311–12) and despite his focus on Walrasian general equilibrium in that work, the microeconomic part of his immensely influential *Economics* (1948) is unmistakeably Marshallian, at any rate as mediated by Chamberlin (1933) and Joan Robinson (1933). And though Marshall had digested Thünen and Cournot, knew the work of Menger and the Austrian School, and admitted that 'there are few writers of modern times who have approached as near to the brilliant originality of Ricardo as Jevons has done' (Marshall 1920, p. 673), yet he 'consistently discounted the "Jevonian revolution"' (Schumpeter 1954, p. 826) and used all his influence, which was great, to insist that in science, as in the world it contemplates, *Natura non facit saltum.* There are few references to Jevons in his famous *Principles,* and in the most extended of these (Appendix I) Marshall went out of his way to counter the former's 'antagonism to Ricardo and Mill' and to defend their value theory against his intemperate exaggerations' (Marshall 1920, pp. 673–6; see also O'Brien 1994, vol. 2, pp. 325–61).

Upon the evidence of Palgrave's original dictionary it appears that by the last decade of the

19th century the effect of Marshall's efforts had been to co-opt Jevons and his 'marginalist' followers into the mainstream of English political economy with a minimum of fuss, and with a minimum of attention to the continental marginalists. Jevons's 'final utility' became 'marginal utility' in Marshall's *Principles* (1920, pp. 78–85), and there was used with deceptive innocence (see Blaug 1996, p. 322–37) to generate a market demand function of price. Though Edgeworth himself contributed 17 articles to the dictionary, including 'Cournot', 'Curves' and 'Demand Curve' in volume 1, 'Mathematical Methods' in volume 2 and 'Pareto', 'Pareto's Law', 'Supply Curve' and 'Utility' in volume 3, his own work was ignored in the general surveys of 'Political Economy' and 'The English School' and his name omitted from the index of volume 1, along with those of Menger and J.B. Clark. Walras received three short references in that volume. Not until volume 3 (1899, pp. 652–5) was his work recognized, and then only for its use of marginal utility. There is no awareness of general equilibrium in that article, and the term appears nowhere else in the original *Dictionary.*

It would appear from the foregoing that if there really was any such thing as a marginal revolution in Anglophone political economy, it began as early as 1767 with Steuart's *Political Œconomy* and still had some way to go by the time volume 3 of the Palgrave dictionary appeared in 1899. Thünen's (1826) generalization of diminishing returns to all factors of production remained unnoticed by any save Marshall. Though Wicksteed (1894) and Flux (1894) reinvented this wheel, Wicksteed's (Palgrave 1899, pp. 140–2) own contribution to the Palgrave article on 'Political Economy' only hints at what later became known as the neoclassical theory of distribution. In 1895 Edgeworth rejected Barone's submission to the *Economic Journal* showing that product exhaustion is implied by Walras's (1894) cost-minimization equations. A companion article to Wicksteed's baldly states that 'the law of DIMINISHING RETURNS points to an increase in the cost of agricultural produce accompanying increase of population' (Palgrave 1899, p. 140). For that author at any rate, nothing had changed since Malthus.

In summary, it would appear that the English School was alive and well in the first decade of the 20th century. Elections to the Political Economy Club included Pigou (in 1906) and J.M. Keynes (1912), along with the Bishop of Stepney (1904), the Rt Hon. Herbert Samuel MP, the Viscount Ridley (1907) and John Buchan (1909). Mill's *Principles* was still perhaps the most widely used textbook. Questions on Adam Smith still appeared in university examinations in political economy (for example, at Edinburgh, 21 November 1898, 17 March 1899). Mathematics was still an unwelcome eccentricity. Jevons (1871, p. vii) had asserted that economics 'must be a mathematical science in matter if not in language'. Marshall (for example, 1890, p. ix) threw all his influence against this doctrine and locked up his own sophisticated mathematics in well-guarded appendices (Keynes 1972, pp. 182–8). Despite his dependence upon mathematical reasoning and his prominence in the emerging profession of economics, Edgeworth's deference for Marshall deterred him from challenging a Cambridge, anti-mathematical orthodoxy that persisted until the 1950s.

Edgeworth was unusual, too, in his ability and willingness to read foreign authors and to recognize their contributions (Keynes 1972, pp. 263–5). In general, the insularity of the English School persisted until well into the 20th century. When Harrod was about to begin his studies in economics, Keynes advised him not to waste his time on the Continent 'where they knew nothing at all of economics' (Harrod 1952, pp. 317–19). The 'market socialists' of the 1930s, none of whom was English, were the first to specify the complete set of marginal conditions required for a welfare optimum in general competitive equilibrium. J.R. Hicks (1939, p. 6) believed himself to be the first English author to 'free the Lausanne School from the reproach of sterility brought against it by the Marshallians'.

It might have been expected that political economists in the United States, at any rate, would have identified with the English School. In the early 19th century authors such as Wayland

(1837) had assimilated Malthus and Ricardo, and as late as 1888 Amasa Walker regarded Jevons and Marshall as 'an extension of the English School' (Goodwin 1972, p. 562). But throughout much of the century protectionist sentiment in the USA was at variance with the ideology of free trade promoted by the English School. And towards the end of that century there was 'an estrangement from British scholarly life' created by a 'growing attachment to German thought' (Goodwin 1972, p. 563). The American Economic Association was originally formed to promote the Liberal-Protestant 'social gospel', very different in spirit and substance from the aristocratic Whiggery of the Political Economy Club.

Bibliography

Anderson, J. 1777. *Observations on the means off exciting a spirit of national industry; chiefly intended to promote the agriculture, commerce, manufactures, and fisheries, of Scotland*. Dublin: Price.

Anon. 1837. 1. *An outline of the science of political economy.* By Nassau W. Senior. London: 1836. 2. *Principes Fondamentaux de l'Économie Politiques, tirés de leçons édites et inédites, de M. N. W. Senior.* Par le Compte Jean Arrivabe. Paris: 1836. *Edinburgh Review* (October), 73–102.

Backhouse, R.E. 2006. Sidgwick, marshall and the Cambridge school of economics. *History of Political Economy* 38: 15–44.

Bacon, R., and W. Eltis. 1976. *Britain's economic problem: Too few producers*. London: Macmillan.

Bagehot, W. 1848. Review of J.S. Mill's principles of political economy. *Prospective Review* 4: 460–502.

Bagehot, W. 1880. *Economic studies,* ed. R.H. Hutton. London: Longmans, Green.

Blaug, M. 1987. Classical economics. In *The New Palgrave: A dictionary of economics*, vol. 1, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Blaug, M. 1996. *Economic theory in retrospect*, 5th ed. Cambridge: Cambridge University Press.

Cairnes, J.E. 1874. *Some leading principles of political economy newly expounded*. London: Macmillan.

Chalmers, T. 1808. *An enquiry into the nature and stability of national resources*. Edinburgh: Moir.

Chamberlin, E.H. 1933. *The theory of monopolistic competition: A reorientation of the theory of value*. Cambridge, MA: Harvard University Press.

Checkland, S.G. 1951. The advent of academic economics in England. *The Manchester School* 19: 43–70.

Cliffe Leslie, T.E. 1879. *Essays in political and moral philosophy.* London: Longmans, Green.

Copleston, E. 1819a. *A letter to the Right Hon. Robert Peel, MP for the University of Oxford, on the pernicious effect of a variable standard of value, especially as it regards the condition of the lower orders and the poor laws . . .*. Oxford: Murray.

Copleston, E. 1819b. *A second letter to the Right Hon. Robert Peel, MP for the University of Oxford, on the causes of the increase in pauperism, and on the poor laws . . .*. Oxford: Murray.

Corsi, P. 1987. The heritage of Dugald Stewart: Oxford philosophy and the method of political economy. *Nuncius* 2: 89–143.

Eltis, W. 2000. *The classical theory of economic growth*, 2nd ed. Basingstoke: Palgrave.

Eltis, S., and W. Eltis. 1997. The life and contribution to economics of the Abbé de Condillac. In E.B. Abbé de Condillac, *Commerce and Government*. Trans. S. Eltis. Cheltenham: Elgar.

Faccarello, G. 1999. *The foundations of Laissez-Faire: The economics of Pierre de Boisguilbert*. London: Routledge.

Fawcett, H. 1863. *Manual of political economy.* London: Macmillan.

Fawcett, M.G. 1870. *Political economy for beginners*. London: Macmillan.

Fetter, F. 1980. *The economist in parliament, 1780–1868*. Durham: Duke University Press.

Feyerabend, P. 1988. *Against method*. London: Verso.

Flux, A.W. 1894. Review of wicksteed. Repr. In *Precursors in mathematical economics*, ed. W. Baumol and S.M. Goldfield. London: London School of Economics, 1968.

Fontana, B. 1985. *Rethinking the politics of a commercial society: The edinburgh review 1802–1832*. Cambridge: Cambridge University Press.

Godwin, W. 1793. *Enquiry concerning political justice and its influence on morals and happiness*. London: Robinson.

Goodwin, C.G.W. 1972. Marginalism moves to the New World. *History of Political Economy* 4: 551–570.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Harrod, R.F. 1952. *The life of John Maynard Keynes*. London: Macmillan.

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

Hollander, J.H. 1932. *Minor papers on the currency question, 1809–1823 by David Ricardo*. Baltimore: Johns Hopkins Press.

Hollander, S. 1987. *Classical economics*. Oxford: Blackwell.

Hollander, S. 1997. *The economics of Thomas Robert Malthus*. 2 vols. Toronto: University of Toronto Press.

Hollander, S. 1998. The canonical classical growth model: Content, adherence and priority. *Journal of the History of Economic Thought* 20: 253–277.

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

Jevons, W.S. 1882. *The state in relation to labour.* London: Macmillan.

Kates, S. 1994. The malthusian origins of the general theory or how Keynes came to write a book about Say's Law and effective demand. *bHistory of Economics Review* 21: 10–20.

Keynes, J.N. 1891. *The scope and method of political economy*. London: Macmillan.

Keynes, J.M. 1972. Essays in biography. In *The collected writings of John Maynard Keynes*, vol. 10, ed. E. Johnson and D. Moggridge. London: Macmillan.

Kojecky, R. 1971. *Eliot's social criticism*. London: Faber.

Lauderdale, J., and E. Maitland. 1804. *An inquiry into the nature and origin of public wealth: And into the means and causes of its increase*. Edinburgh: Constable.

Malthus, T.R. 1798. *An essay on the principle of population as it affects the future improvement of society, with remarks upon the speculations of Mr Godwin, M. Condorcet, and other writers*. London: Johnson.

Malthus, T.R. 1800. *An investigation of the cause of the present high price of provisions. By the author of the essay on the principle of population*. London: Johnson.

Malthus, T.R. 1803. *An essay on the principle of population, or, A view of its past and present effects on human happiness, with an inquiry into our prospects respecting the future removal or mitigation of the Evils which it occasions*. London: Johnson.

Malthus, T.R. 1815a. *An inquiry into the nature and progress of rent, and the principles by which is regulated*. London: Murray.

Malthus, T.R. 1815b. *The grounds of an opinion on the policy of restricting the importation of Foreign Corn . . .*. London: Murray.

Malthus, T.R. 1817. *An essay on the principle of population . . . 4th edn of Malthus (1803), described as 5th*. London: Hunter.

Malthus, T.R. 1820. *Principles of political economy, considered with a view to their practical application*. London: Pickering.

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

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan. 1952.

Marx, K. 1873. *Capital: A critique of political economy,* ed. F. Engels, Trans. S. Moore and E. Aveling. 3 vols. Moscow: Progress Publishers, 1954.

McCulloch, J.R. 1843. *Principles of political economy.* Edinburgh: Tait.

McCulloch, J.R. 1845. *Literature of political economy: A classified catalogue*. London: Longmans.

Mill, J. 1821. *Elements of political economy.* London: Baldwin et al.

Mill, J.S. 1869. Thornton on labour and its claims. In *Collected works of John Stuart Mill*, vol. 5, ed. J.M. Robson. Toronto: University of Toronto Press, 1967.

Mill, J.S. 1871. *Principles of political economy,* 7th edn. ed. W.J. Ashley. London: Longmans, Green, 1909.

Mrs Marcet (Jane Haldimand). 1817. *Conversations on political economy: In which the elements of that science are familiarly explained.* London: Longman et al.

Nicholson, J.S. 1893–1901. *Principles of political economy.* 3 vols. London: Macmillan.

O'Brien, D.P. 1994. *Methodology, money and the firm.* 2 vols. Aldershot: Elgar.

O'Brien, D.P. 2004. *The Classical economists revisited*. Princeton: Princeton University Press.

Paley, W. 1802. *Natural theology*. London: Wilkes and Taylor.

Palgrave, R.H. 1894–1899. *Dictionary of political economy.* 1899. Vol. 1, 1894; Vol. 2, 1896; Vol. 3. London: Macmillan.

Peart, S.J., and D.M. Levy. 2003. 1830–1870: Post-Ricardian British economics. In *A companion to the history of economic thought*, ed. W.J. Samuels, J.E. Biddle, and J.B. Davis. Oxford: Blackwell.

Political Economy Club. 1921. *Minutes of proceedings, 1899–1920, roll of members and questions discussed, 1821–1920, with documents bearing on the history of the club*. London: Macmillan.

Pryme, G. 1823. *Introductory lecture and syllabus*. Cambridge: Cambridge University Press.

Ricardo, D. 1815. *An essay on the influence of a low price of corn on the profits of stock*. London: Murray.

Ricardo, D. 1817. *On the principles of political economy and Taxation*. London: Murray.

Ricardo, D. 1951–73. *The works and correspondence of David Ricardo*. 11 vols, ed. P. Sraffa. Cambridge: Cambridge University Press.

Robinson, J.V. 1933. *The economics of imperfect competition*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Samuelson, P.A. 1948. *Economics: An introductory analysis*. New York: McGraw Hill.

Samuelson, P.A. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16: 1415–1434.

Say, J.-B. 1803. *Traitè d'economie politique: ou, Simple exposition de la manière dont se forment, se distribuent et se consomment les richesses*. Paris: Dèterville.

Schumpeter, J.A. 1954. *History of economic analysis*. London: Allen & Unwin.

Senior, N.W. 1836. *An outline of the science of political economy*. London: W. Clowes.

Sidgwick, H. 1883. *The principles of political economy.* London: Macmillan.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. 2 vols., ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Oxford University Press, 1976.

Smith, V.E. 1956. Malthus's theory of demand and its influence on value theory. *Scottish Journal of Political Economy* 3: 205–220.

Sraffa, P. 1926. The laws of return under competitive conditions. *Economic Journal* 35: 535–550.

Steuart, J. 1767. *An inquiry into the principles of political Œconomy: Being an essay on the science of domestic policy in free nations. In which are particularly considered population, agriculture, trade, industry, money,*

coin, interest, circulation, banks, exchange, public credit, and taxes. 2 vols. London: Millar and Cadell.
Stewart, D. 1792, 1814/1827. *Elements of the Philosophy of the Human Mind,* 3 vols. Reprinted in *The collected works of Dugald Stewart*, 11 vols. ed. W. Hamilton. Edinburgh: Constable, 1854–60.
Stigler, G. 1952. The Ricardian theory of value and distribution. *Journal of Political Economy* 60: 187–207.
Stigler, G. 1958. Ricardo and the 93 % labor theory of value. *American Economic Review* 48: 357–367.
Sumner, J.B. 1816. *A treatise on the records of the creation: With particular reference to Jewish History, and the consistency of the principle of population with the wisdom and goodness of the deity.* 2 vols. London: Hatchard.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*. London: Hatchard.
Thornton, R.W.T. 1869. *On labour: Its wrongful claims and rightful dues, its actual present and possible future*. London: Macmillan.
Torrens, R. 1815. *An essay on the external corn trade*. London: Longman et al.
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.* London: Longman et al.
Turgot, A.R.J. 1768. Observations sur le mémoire de M. de Saint-Péravy en faveur de l'impôt indirect. In *Écrits Économiques*, ed. B. Cazes. Paris: Calman-Lévey, 1970.
von Thünen, J.H. 1826. *Der isolirte Staat in Beziehung auf Landwirthschaft und Nationalökonomie, part 1*. Hamburg: Pethes.
Walras, L. 1894. *Élèments d'èconomie politique pure: ou, Thèorie de la richesse sociale*. Lausanne: Rouge.
Walsh, V., and H. Gram. 1980. *Classical and neoclassical theories of general equilibrium*. New York: Oxford University Press.
Waterman, A.M.C. 1991a. *Revolution, economics and religion: Christian political economy, 1798–1833*. Cambridge: Cambridge University Press.
Waterman, A.M.C. 1991b. The 'canonical classical model' in 1808 as viewed from 1825: Thomas Chalmers on the national resources. *History of Political Economy* 23: 221–241.
Waterman, A.M.C. 1995. Why William Paley was 'The first of the Cambridge Economists'. *Cambridge Journal of Economics* 20: 673–686.
Waterman, A.M.C. 1998. Reappraisal of 'Malthus the economist', 1933–97. *History of Political Economy* 30: 293–334.
Waterman, A.M.C. 1999. Hollander on the 'canonical classical growth model': A comment. *Journal of the History of Economic Thought* 21: 311–313.
Waterman, A.M.C. 2004. *Political economy and christian theology since the enlightenment*. Basingstoke: Palgrave Macmillan.
Wayland, Francis. 1837. *The elements of political economy.* Boston: Gould and Lincoln.
West, E. 1815. *Essay on the application of capital to land; with observations shewing the impolicy of any great restriction on the importation of corn . . . by a fellow of University College, Oxford*. London: Underwood.
Whately, R. 1831. *Introductory lectures in political economy*. London: Fellowes.
Wicksteed, P.H. 1894. *Essay on the coordination of the laws of distribution*. London: Macmillan.
Winch, D.N. 1983. Science and the legislator: Adam Smith and after. *Economic Journal* 93: 501–520.
Winch, D.N. 1996. *Riches and poverty. An intellectual history of political economy in Britain, 1750–1834*. Cambridge: Cambridge University Press.
Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Enlightenment, Scottish

John Robertson

Abstract

The Scottish contribution to the Europe-wide intellectual movement of Enlightenment in the 18th century was unusually rich, covering moral philosophy, history, and political economy. It was not the simple product of the Union with England in 1707; more important were the gradual opening up of intellectual life and reform of the country's intellectual institutions, notably the universities, and economic growth, rapid by the last quarter of the century. The Scots set the investigation of economic phenomena in a broad framework; led by David Hume and Adam Smith, they were particularly interested in the comparative development prospects of rich and poor nations.

Keywords

Chalmers, T.; Commercial society; Division of labour; Enlightenment; Ferguson, A.; Galiani, F.; Genovesi, A.; Hume, D.; Hutcheson, F.; Individual liberty; Innovation; Invisible hand; Justice; Kames, Lord; Law, J.; Luxury; Mandeville, B.; Melon, J.-F.; Millar, J.; Natural jurisprudence; Physiocracy; Political economy; Private property; Protectionism;

Pufendorf, S.; Quesnay, F.; Reid, T.; Robertson, W.; Rule of law; Scottish Enlightenment; Shaftesbury, Lord; Smith, A.; Specialization; Stages theory of development; Steuart, Sir J.; Stewart, D.; Subsistence; *Tableau économique*

JEL Classifications
B1

Between 1740 and 1790 Scotland provided one of the most distinguished branches of the European Enlightenment. David Hume and Adam Smith were the pre-eminent figures in this burst of intellectual activity; and around them clustered a galaxy of major thinkers, including Francis Hutcheson, Lord Kames, Adam Ferguson, William Robertson, Thomas Reid, Sir James Steuart and John Millar. The interests of individual thinkers ranged from metaphysics to the natural sciences; but the distinctive achievements of the Scottish Enlightenment as a whole lay in those fields associated with the enquiry into 'the progress of society' – history, moral and political philosophy and, not least, political economy.

'Enlightenment' and 'Scottish Enlightenment' were usages unknown in the 18th century: the term 'Scottish Enlightenment' was first coined in the early 20th century, and began to be generally used by historians in the 1960s. (*Lumières* and *Aufklärung* were in 18th-century use, but not to denote a European Enlightenment as a whole.) As a historian's construction, however, the term 'Scottish Enlightenment' is supported by the consciousness of those named above that they shared common intellectual interests (which did not preclude disagreement between them) and a common standing as men of letters in 18th-century Scottish society. This awareness of belonging to a broad intellectual movement extended to the continent of Europe: led by Hume, the Scottish thinkers cultivated connections with Paris, the Enlightenment's acknowledged metropolitan centre. But the Scottish Enlightenment is perhaps best understood when it is compared with the Enlightenment in Italy or in Germany. The concern with economic improvement and its moral and political conditions and consequences was as urgent, for instance, in the distant Kingdom of Naples as in Scotland; and political economy was equally absorbing to the Neapolitan philosophers Antonio Genovesi and Ferdinando Galiani.

At the same time, the experience of Scotland in the 18th century was distinctive in a number of respects, which offered a particular stimulus to Scottish thinkers. First of all, there was the actual achievement of economic growth. The late 17th-century Scottish economy supported an uneasy balance between population and food supply; bad harvests, which occurred in a sequence in the 1690s, could cause severe shortages and even localized famine. Overseas trade was likewise vulnerable. Nevertheless the elites, both landed and urban, were committed to economic development, and showed a marked propensity to invest. Agriculture gradually became commercialized, and landowners joined merchants to invest in manufactures, and, most spectacularly, in the 'Darien venture', intended to establish a Scottish trading colony in Panama. The failure of the latter persuaded many of the elite that economic development could only come through closer union with England. In the event, the economic fruits of the Union were disappointingly slow in coming; but by the third quarter of the 18th century it was clear to contemporaries that agriculture, trade and manufactures were all on an upward curve. The thinkers of the Scottish Enlightenment thus enjoyed an unusually direct acquaintance with the phenomena of economic development.

Scotland's political position was also unusual. Many of Europe's monarchies sought to bring their constituent kingdoms into closer union over the 18th century, for economic as well as administrative reasons. But none did so as successfully as the British monarchy. The Union of 1707 with England was in no simple or direct sense the cause of Scotland's economic growth (or of its Enlightenment). But it secured a common framework of law and a common market, and it also established that the Scottish Presbyterian and the English Anglican Churches should coexist in peace. These gains were important to the great majority of the Scottish elites, and it was never in their

interest to back the Jacobite challenge to the Hanoverian monarchy.

Culturally and intellectually, the position of Scotland looked unpropitious before 1700. There were pockets of interest in the new science, Newton having a group of Scottish adherents; but the latest developments in French philosophy were shunned for their Epicurean, materialist and sceptical tendencies. After the Revolution of 1688, however, change gradually got under way in the institutions most important for intellectual life, making possible the infiltration of new ideas. The fierce, covenanting Presbyterianism of the 17th century was dissipated, as the 'Moderate' group of clergy rose to power in the Kirk. The universities of Edinburgh, Glasgow, Aberdeen and St Andrews were reformed, allowing professorial specialization; and around the universities there developed a vigorous informal culture of voluntary clubs, most famous of which was the Select Society of Edinburgh, founded by David Hume and his friends in 1754. Together these changes secured for Scottish thinkers unprecedented intellectual freedom and social support; and they provided an object lesson in the importance of the moral and cultural as well as the material dimensions of progress.

The intellectual interests which distinguished the Scottish Enlightenment had two more specific sources. One was the explicit preoccupation with the conditions and means of economic development which was fostered by the debate which preceded the Union of 1707. The preoccupation was by no means unique to the Scots, but the contributions of John Law (the future author of the French Mississippi Scheme) and others ensured a high quality of discussion. The other, two decades later, was the initiative taken by two very different philosophers, Francis Hutcheson and David Hume, to transform the agenda by which philosophy was taught and discussed in Scotland. Drawing on the moral philosophy of Shaftesbury and the natural jurisprudence of Pufendorf, Hutcheson taught his Glasgow students, who included Adam Smith, a moderate, benevolent, providential Stoicism. More disturbingly, Hume drew on the scepticism of Pierre Bayle and the Epicurean morals of Bernard Mandeville to offer in his *Treatise of Human Nature* (1739–40) and his two later *Enquiries* (1748; 1751) an account of justice and morals which had no need of divine support. Most of those now associated with the Scottish Enlightenment found Hutcheson's philosophy more congenial; but it was Hume's challenge which galvanized them. It was Hume, moreover, who turned their attention back to economic matters. Recognizing that philosophy alone would never make the Scots into virtuous atheists, Hume decided instead to educate them in political economy, the subject of the leading essays in his *Political Discourses* of 1752.

For Hume as for all the Scottish thinkers, political economy was not a science apart. It belonged within a wider enquiry into the 'progress of society'. There were three principal dimensions to this enquiry: the historical, the moral and the political.

The historical theory of the Scottish Enlightenment developed a line of argument from later 17th-century natural jurisprudence, a tradition made familiar to the Scots by its incorporation in the moral philosophy curriculum of the reformed universities. Discarding the older jurisprudential thesis of the contractual foundations of society and government, the Scots focused on the new insights of Pufendorf and Locke into the origin and development of property. According to Pufendorf, there had never been an original state of common ownership of land and goods; from the first, property was the result of individual appropriation. As increasing numbers made goods scarce, individual property became the norm, and systems of justice and government were established to secure it. What the Scots added to this argument was a scheme of specific stages of social development, the hunting, the pastoral, the agricultural and the commercial. At each of the four stages the extent of property ownership was related to the society's means of subsistence, and these shaped the nature and sophistication of the society's government. Different versions of the theory were offered by Adam Ferguson in his *Essay on the History of Civil Society* (1767) and by John Millar in his *Origin of the Distinction of Ranks* (1770), and it underlay both Lord Kames's investigations into

legal history and William Robertson's historical narratives. The locus classicus of the theory, however, was Adam Smith's Lectures on Jurisprudence, delivered to his students in Glasgow in the early 1760s.

As Smith's exposition makes particularly clear, the stages theory of social development provided the historical premises for political economy. An explicitly conjectural theory – a model of society's 'natural' progress – it provided a framework for a comparably theoretical treatment of economic development as 'the natural progress of opulence'. By positing the systematic interrelation of economic activity, property and government, with consequences which could be neither foreseen nor controlled by individuals, the theory also underlined the limits of effective government action. 'Reason of state', the standby of rulers and their advisers for over two centuries, still had the capacity to distort and obstruct the economic activity of subjects and those with whom they would trade; but the Scots' historical perspective showed it to be a doctrine inadequate to the complexity of a modern commercial economy.

The moral thought of the Scottish Enlightenment was closely related to the historical, sharing a common origin in 17th-century natural jurisprudence. Here the inspiration was the jurisprudential thinkers' increasingly sophisticated treatment of needs. These, it was recognized, could no longer be thought of primarily in relation to subsistence; with the progress of society, needs must be understood to cover a much wider range of scarce goods, luxuries as well as necessities. The potential of this insight was seen by every Scottish moral philosopher, but again it was Smith who exploited it to the full, in the *Theory of Moral Sentiments* (1759). Beyond the most basic necessities, Smith acknowledged, men's needs were always relative, a matter of status and emulation, of bettering one's individual condition. But it was precisely the vain desires of the rich and the envy of others which served, by 'an invisible hand', to stimulate men's industry and hence to increase the stock of goods available for all ranks.

Such an argument, however, had to overcome two of the most deeply entrenched convictions of European moral thought: the Aristotelian view that the distribution of goods was a matter for justice, and the classical or civic humanist view that luxury led to corruption and the loss of moral virtue. The Scots answered the first more confidently (but perhaps less satisfactorily) than the second. Following Grotius, Hobbes and Pufendorf, they defined justice in exclusively corrective terms, setting aside questions of distribution. On the issue of corruption, they were divided. Hume, who ridiculed fears of luxury, was the most confident; Ferguson, who defiantly reasserted the ancient ideal of virtue, was the most pessimistic. Smith was closer to Hume in preferring propriety to virtue, at least for the great majority; but he showed that he shared Ferguson's doubts when he added, at the end of his life, that the disposition to admire the rich and the great did tend to corrupt moral sentiments. At a fundamental level, however, there was general agreement. As a consequence of the progress of society, the multiplication of needs was not only irreversible; it was the essential characteristic of a 'cultivated' or 'civilized' as distinct from a 'barbarian' society. And civilization, however morally ambiguous, was preferable to barbarism. With consensus on this, the moral premises of political economy were secure.

The definition of justice in simple corrective terms provided the starting-point for the political dimension of the Scottish enquiry. The priority of any government, the Scots believed, must be the security of life and property, ensuring every individual liberty under the law. This, as Smith put it, was freedom 'in our present sense of the word'; and there was a general confidence that it was tolerably secure under the governments of modern Europe, including the absolute monarchies. In principle, individual liberty was a condition of a fully commercial society: its provision, therefore, was the institutional premise of political economy.

Few of the Scots took their analysis beyond this relatively simple, if vital, point; the theory of the modern commercial state was not a Scottish achievement. Both Hume and Smith were more concerned to limit the opportunities for enlarging government at the expense of 'productive' society, by confining the former to the minimum necessary provision of justice, defence and public works.

But they also recognized that the proliferation of interests in a commercial society would require more sophisticated institutional mechanisms to ensure their adequate representation within the political system. Smith's analysis in Book IV of the *Wealth of Nations* of the growing alienation of the colonial elites in North America from parliamentary authority was an object lesson in the need for such representation – and a strong hint that it was incompatible with maintaining an extended empire.

A large part of the originality of the Scottish Enlightenment's conception of political economy lay in this exploration of the historical, moral and institutional framework of economic activity. But of course the Scots also engaged directly in economic analysis; and one such work of analysis, Adam Smith's *Wealth of Nations* (1776), would so outshine all others that it came to be regarded as having established political economy as a science in its own right.

The Scots' attention focused on growth in a context of international rivalry. In contemporary terms, Hont has shown, the issue was the means by which poor countries (of which Scotland might be regarded as one) could best hope to catch up on rich countries (such as England certainly was). What is striking is the hard- headedness with which Hume and Smith tackled the issue. Responding to French economists – Hume to Jean-François Melon, Smith to the Physiocrats – who argued that agriculturally endowed countries should follow a different path from purely commercial nations, the Scots insisted that one analysis applied to all. Protection for agricultural economies and their manufactures, a policy supported by the former Jacobite exile Sir James Steuart in his *Principles of Political Economy* (1767), was futile and damaging. But theirs was no naive optimism in the equalizing powers of commerce. The ideal of *doux commerce*, by which trade would be the agent of global peace and prosperity, was as much of a panacea as the belief that commercial success would be self-cancelling, because the advantage of low labour costs would always pass on to others. Instead, Hume and Smith suggested that rich countries could expect to maintain their advantage over poorer ones, whether by flexible specialization and product innovation (Hume) or by constantly increasing industrial productivity through the division of labour (Smith). What distinguished commercial superiority from military conquest was that it was achieved 'without malice'; poor countries would also develop if they followed the same route, even if they might never catch up on the rich.

Brilliant as Hume's economic essays were, it was Adam Smith's *Wealth of Nations* (1776) which set the standard of Enlightenment political economy. To be systematic and comprehensive had earlier been the ambition, at least, of Quesnay's *Tableau Economique* (1758–9), Genovesi's *Lezioni di Commercio* (1765) and Steuart's *Principles;* but the *Wealth of Nations* eclipsed them all. Its success, moreover, was such as to suggest that political economy had an identity all of its own. Smith himself did not admit such an implication, continuing to insist that political economy was but 'a branch of the science of a statesman or legislator': his own engagement with both jurisprudence and moral philosophy left him disinclined to drop the wider intellectual framework in which political economy had been conceived. But a work at once as extensive and as self-contained as the *Wealth of Nations* made it at least plausible to suppose that what it presented was a distinct, autonomous science of political economy.

Smith's death in 1790 coincided with the end of the Scottish Enlightenment. In Scotland as throughout Europe, the French Revolution transformed the conditions and assumptions of intellectual life, while political economy had to come to terms with the increasingly obvious impact of machinery. Within Scotland Dugald Stewart set himself to adapt the Enlightenment conception of political economy to these new circumstances; but while he had French admirers, his expansive, didactic approach had few followers in Britain. Another Scot, Thomas Chalmers, took the lead alongside Malthus in attaching political economy to newly urgent theological concerns, while Ricardo and his followers simply took a narrower view of the subject. Even so, it would be a mistake to see 19th-century classical political economy as

a new departure. As the philosophical analysis of Hegel (who learnt much from Steuart) and the radical critiques of Marx and the early socialists pointed out, the historical, moral and institutional premises on which political economy rested were still those elucidated by the Scots.

See Also

- ▶ Hume, David (1711–1776)
- ▶ Hutcheson, Francis (1694–1746)
- ▶ Mandeville, Bernard (1670–1733)
- ▶ Pufendorf, Samuel von (1632–1694)
- ▶ Smith, Adam (1723–1790)
- ▶ Steuart, Sir James (1713–1780)
- ▶ Stewart, Dugald (1753–1828)

Bibliography

Berry, C.J. 1997. *Social theory and the Scottish Enlightenment*. Edinburgh: Edinburgh University Press.

Broadie, A. (ed.). 2003. *The Cambridge companion to the Scottish Enlightenment*. Cambridge: Cambridge University Press.

Devine, T.M. 1994. *The transformation of rural Scotland: Social change and the agrarian economy 1660–1815*. Edinburgh: Edinburgh University Press.

Hont, I. 2005. *Jealousy of trade: International competition and the nation-state in historical perspective*. Cambridge, MA: Harvard University Press.

Hont, I., and M. Ignatieff (eds.). 1983. *Wealth and virtue: The shaping of political economy in the Scottish Enlightenment*. Cambridge: Cambridge University Press.

Phillipson, N.T. 1981. The Scottish Enlightenment. In *The Enlightenment in national context*, ed. R. Porter and M. Teich. Cambridge: Cambridge University Press.

Robertson, J.C. 2005. *The case for the Enlightenment. Scotland and Naples 1680–1760*. Cambridge: Cambridge University Press.

Sakamoto, T., and H. Tanaka (eds.). 2003. *The rise of political economy in the Scottish Enlightenment*. London: Routledge.

Sher, R.B. 1985. *Church and university in the Scottish Enlightenment*. Princeton/Edinburgh: Princeton University Press; Edinburgh: Edinburgh University Press.

Sher, R.B. 2006. *The Enlightenment and the book. Scottish authors and their publishers in eighteenth-century Britain, Ireland and America*. Chicago/London: Chicago University Press.

Winch, D. 1996. *Riches and poverty: An intellectual history of political economy in Britain 1750–1834*. Cambridge: Cambridge University Press.

Enterprise Zones

Leslie E. Papke

Abstract

Enterprise zones are geographically targeted economic development incentives used in the United States by individual states since the early 1980s and the federal government since 1993. Research on state zone programmes that accounts for the endogeneity of zone designation finds little improvement in the employment and incomes of zone residents, but some evidence that firms respond to tax incentives for capital. In contrast, the federal empowerment zone programme combines tax incentives with local initiatives and access to large federal grants. Recent research on round one of the federal programme finds mixed evidence on zone resident employment.

Keywords

Enterprise Zones; Investment Subsidies; Labour Subsidies; Tax Competition; Tax Credits; Tax Incentives

JEL Classifications

H3

Enterprise zone programmes are geographically targeted tax, expenditure, and regulatory inducements used by US state and local governments since the early 1980s and by the federal government since 1993. While they differ in their specifics, all the programmes provide development incentives, including tax preferences to capital and/or labour, in an attempt to induce private investment location or expansion to depressed areas and to enhance employment opportunities for zone residents. Most enterprise zones are designated in urban areas, but there are some rural zones. Typically, state and local zone programmes provide larger tax credits for business investment than for employment incentives. Investment

incentives include the exemption of business-related purchases from state sales and use taxes, investment tax credits and corporate income or unemployment tax rebates. Labour subsidies include employer tax credits for all new hires or zone-resident new hires, employee income tax credits and job-training tax credits. Some programmes assist firms financially with investment funds or industrial development bonds.

Enterprise zones have been criticized as ineffective and inefficient in stimulating new economic activity. This criticism is part of a long-standing debate on the effects of intersite tax differentials on the location of capital investment. It is argued that if tax-induced investment represents only relocation from another state, then tax competition is a zero-sum game for the country as a whole. In addition, the preferential treatment of certain types of investment or employment within enterprise zones may induce decisions that would not be economically sound in the absence of the tax incentives. Often, however, redistribution of economic activity within a state may be a desirable goal. If investment is relocated from local labour markets with low unemployment to local labour markets with higher unemployment, the incentives may generate efficiency gains for the economy as underutilized resources are tapped (Bartik 1991). Efficiency gains may also result if reductions in unemployment produce positive externalities, such as reductions in social unrest.

A partial equilibrium model predicts that a labour subsidy or an equal-cost subsidy to both zone capital and zone resident labour will raise zone wages. A capital subsidy alone may actually reduce zone wages – yet many of the subsidies are for capital investment in the zone (Gravelle 1992; Papke 1994).

Empirical evaluations of zone programmes typically measure the amount of investment undertaken after the designation, for example, or the increase in the number of firms in the zone, and the change in zone employment. Two key methodological issues in empirical evaluations are (*a*) to separate the effects of zone designation from jobs and investments arising from other factors – for example, general upswings in the economy; (*b*) to account for the depressed economic characteristics that led to the initial zone designation. If zone sites are better randomly selected, the effect of the programme can be measured by comparing the performance of the experimental and control groups. But zone designation in the 43 state and local programmes in the United States depends on comparative unemployment rates, population levels and trends, poverty status, median incomes, and percentage of welfare recipients, so the data are non-experimental. This sample selection problem can be addressed with a variety of econometric techniques.

Econometric analysis of a zone's success faces a practical difficulty in that conventional economic data are not available by zone. In most states, zones do not coincide with census tracts or taxing jurisdictions. As a result, zone areas cannot be pinpointed in standard data collections. Zip code level data is available from the Census, but outcome measures are ten years apart.

Econometric evaluations of the Indiana and New Jersey programmes find mixed effects on investment and employment. Indiana zones are estimated to have greater inventory growth and fewer unemployment claims than they would have in the absence of the zone designation (from 1983 to 2006, an inventory tax credit was the most lucrative incentive). However, in the 1980s, inventory investment came at the cost of a drop in the value of depreciable property (Papke 1994). Moreover, despite the reduction in unemployment rates in the zones, a comparison of incomes from the 1980 and 1990 Censuses suggests that zone residents are not appreciably better off after the first decade of the Indiana zone programme (Papke 1993) and there is no discernable increase in capital investment or land values (Papke 2001). Similar econometric analysis of the New Jersey enterprise zone programme finds no positive effects on either business investment or employment (Boarnet and Bogart 1996). Multi-state econometric analyses that combine data from many states – thereby assuming zone programmes have similar effects in every state – typically find no positive zone effects on business activity or employment (Bondonio 2003; Bondonio and Engberg 2000). Peters and Fisher (2002) survey state evaluations.

Cost-per-job estimates from zone programmes are rare. The literature also lacks a discussion of the distribution of the cost of the zone programme between state and local governments. For example, local governments may bear the brunt of the cost of a state enterprise zone programme if tax incentives are provided against local taxes without state reimbursement.

Congress established the Empowerment Zone and Enterprise Community (EZ/EC) programme in 1993 and the Renewal Community (RC) programme in 2000 to provide assistance to the nation's distressed communities. By 2007, there had been three rounds of EZs, two rounds of ECs, and one round of RCs leading to a total of 40 empowerment zones (30 urban and 10 rural), 95 enterprise communities (65 urban, 30 rural) and 40 renewal communities.

Empowerment zone incentives include a 20 per cent employer wage credit for the first 15,000 dollars of wages for zone residents who work in the zone, additional expensing of equipment investments of qualified zone businesses, and expanded tax exempt financing for certain zone facilities. Each zone is eligible for 100 million dollars in Social Services Block Grant funds. Selected areas needed to demonstrate pervasive poverty, unemployment and general distress, and applicants had to outline a plan of action that included local business and community interests. The residence- based approach of the income tax credit differs significantly from another federal programme designed to increase employment of the disadvantaged. The Targeted Jobs Tax Credit provides firms with a similar-sized subsidy for wages paid to targeted individuals – primarily welfare recipients and poor youth. Providing a subsidy based on individual characteristics may create a stigma that actually reduces the probability of being hired. Residence-based eligibility may eliminate this problem and encourage individuals who become employed to continue to live in the zone.

Features of the programmes have changed over time. Round I and II EZs and ECs received different combinations of grant funding and tax benefits. By round III, EZs and the RCs received mainly tax benefits. The GAO (1991, 2004, 2006) reports that Round I and II EZs and ECs are continuing to access their grant funds and Internal Revenue Service (IRS) data show that businesses are claiming some tax benefits (Brashares 2000). However, the IRS does not collect data on other tax benefits and cannot always identify the communities in which they were used. The lack of tax benefit data limits evaluation of the programmes.

Evaluation of the federal programme is also confounded by its hybrid structure. The federal EZ/EC programme is based on the idea that effective community revitalization results when the strategy is tailored to the local site. The diverse nature of the Round I EZ/ECs – each may differ in terms of objective, size of targeted area, type of designation, governance structure, projects used, grant money, and strategies for implementation – has made it difficult to generate general conclusions about even the early stages of Round I implementation (GAO 2004, 2006). Further, the tax incentives changed over the three rounds of the federal programme. Third, no easy method of data collection was included in the tax forms so even usage is hard to measure.

Using Census data, Hanson (2007) finds no effect of the first round zone programme on local employment or poverty rates in the targeted areas, but instead finds capitalization into property values. Busso and Kline (2006) find modest improvements in labour market conditions, but sizable increases in owner-occupied housing values and rents along with small changes in the demographic composition of neighbourhoods. Taken together, these two papers suggest that improvements for residents have been limited at best, but that property owners have benefited from the federal programme.

See Also

- ▶ Economic Development and the Environment
- ▶ Fiscal Federalism
- ▶ Local Public Finance
- ▶ Public Finance
- ▶ Regional Development, Geography of
- ▶ Taxation and Poverty

Bibliography

Bartik, T.J. 1991. *Who benefits from state and local economic development policies?* Kalamazoo: W.E. Upjohn Institute for Employment Research.

Boarnet, M.G., and W.T. Bogart. 1996. Enterprise zones and employment: Evidence from New Jersey. *Journal of Urban Economics* 40: 198–215.

Bondonio, D., and J. Engberg. 2000. Enterprise zones and local employment: Evidence from the states' programs. *Regional Science and Urban Economics* 30: 519–549.

Brashares, E. 2000. Empowerment zone tax incentive use: What the 1996 data indicate. *Statistics of Income Bulletin.*

Busso, M., and P. Kline. 2006. *Do local economic development programs work? Evidence from the federal empowerment zone program*. Mimeo: University of Michigan.

Engberg, J., and R. Greenbaum. 1999. State enterprise zones and local housing markets. *Journal of Housing Research* 10: 163–187.

GAO (General Accounting Office). 1991. *Businesses' use of empowerment zone incentives. RCED-99–253*. US Government Accounting Office: Washington, DC.

GAO. 2004. *Community development: Federal revitalization programs are being implemented, but data on the use of tax programs are limited. RCED 04–306*. Washington, DC: US Government Accounting Office.

GAO. 2006. *Empowerment zone and enterprise community program: Improvements occurred in communities but the effect of the program is unclear. RCED-06–727*. Washington, DC: US Government Accounting Office.

Gravelle, J.G. 1992. *Enterprise zones: The design of tax incentives*, CRS Report for Congress 92–476 S. Washington, DC: Congressional Research Service, Library of Congress.

Hanson, A. 2007. *Poverty reduction and local employment effects of geographically targeted tax incentives: An instrumental variables approach*. Mimeo: Syracuse University.

HUD (U.S. Department of Housing and Urban Development). 1992. *State enterprise zone update: Summaries of the state enterprise zone programs*. Washington, DC: U.S. Department of Housing and Urban Development.

Papke, L.E. 1993. What do we know about enterprise zones? In *Tax policy and the economy*, ed. J.M. Poterba, vol. 7. Cambridge, MA: MIT Press.

Papke, L.E. 1994. Tax policy and urban development: Evidence from the Indiana enterprise zone program. *Journal of Public Economics* 54: 37–49.

Papke, L.E. 2001. *The Indiana enterprise zone revisited: Effects on capital investment and land values. National Tax Association Proceedings of the Ninety-Third Annual Conference*. Washington, DC: National Tax Association.

Peters, A.H., and P.S. Fisher. 2002. *State enterprise zone programs: Have they worked?* Kalamazoo: W.E. Upjohn Institute for Employment Research.

Entitlements

Hillel Steiner

In the strong sense, an entitlement is something owed by one set of persons to another. The thing owed is either a performance of a certain kind, such as a dental extraction, or a forbearance from interfering from some aspect of the title-holder's activity or enjoyment, such as not trespassing on someone's land. Strong entitlements imply the presence of a right in the person entitled and a corresponding or *correlative* obligation in the person owing the performance or forbearance. Typically, the person entitled is further vested with ancillary powers to waive the obligation or, alternatively, to initiate proceedings for its enforcement. A secondary (and contested) instance of a strong entitlement arises with respect to the position of a third-party beneficiary of a right-obligation relation between two other parties, such as the beneficiary of an insurance policy. Third parties usually lack powers of waiver and enforcement, for it is not strictly to them that fulfilment of the obligation is owed.

A weaker form of entitlement may be said to pertain to those of a person's activities which, while not specifically protected by obligations in others not to interfere, are nevertheless indirectly and extensively protected by their other forbearance obligations. Thus, while persons may be under no obligation specifically to allow someone to use a pay telephone, they probably do have forbearance obligations with respect to assault, theft, property damage, etc., the joint effect of which is to afford some high (but incomplete) degree of protection to someone using a pay telephone. However, such an entitlement amounts to less than the full protection afforded by a right inasmuch as it does not, for example, avail against anyone who may already be using that telephone.

Beyond strong and weak entitlements, one may also possess many largely unprotected liberties. These consist in those activities from which one has no obligation to refrain but with which,

equally, no direct or extensive indirect claims to non-interference. So, broadly speaking, persons' strong entitlements may be construed as conjunctively constituting their spheres of ownership, while their weak entitlements and their unprotected liberties constitute the fields of activity within which they exercise the powers and privileges of ownership. Normally, it is persons' strong entitlements that are of primary normative concern, with weak entitlements and unprotected liberties being determined residually.

Entitlements may be either legal or moral. Sets of legal entitlements tend to reflect the multifarious demands of various customs, moral principles, judicial decisions and state policy. A set of moral entitlements, on the other hand, is commonly derived from some basic principle embedded in a moral code. The nature of this derivation varies with the type of code involved. In many single-value codes (such as utilitarianism), entitlements are instrumental in character: whether and what sort of an obligation is owed, by one person to another, depends upon the relative magnitude of the contribution that fulfilment of that obligation would make to realizing that value. Changing causal conditions of maximization warrant alterations in the content and distribution of entitlements. Codes containing a plurality of independent values characteristically generate entitlements from a principle of justice. The set of entitlements thus derived possesses intrinsic and not merely instrumental value, though its normative status depends upon the ranking of justice in relation to the code's other values. In such codes, the chief distinction between moral obligations that (like kindness) are not correlative to any entitlement and those of justice that are, lies in the fact that only the latter are waivable and permissibly enforceable.

Much of the philosophical treatment of entitlements is located in discussions of rival theories of justice. These theories differ according to the various norms they propose for determining who owes what to whom. Endorsing the classical formal conditions of justice – 'rendering to each what is due to him' and 'treating like cases alike' – they diverge widely in their interpretations of what is due to a person and what count as like cases. Procedural and substantive criteria that have been offered for determining individuals' entitlements include: relative need, productivity, equal freedom, equal utility, personal moral worth, interpersonal neutrality, personal inviolability, initial contract and so forth. As is immediately obvious, the nature and distribution of the entitlements mandated by each of these criteria are by no means self-evident, and their identification thus requires supplementary postulates that are variously drawn from psychological theories, from theories of moral and rational choice, and from conceptual analyses of the criteria themselves. It is also true that not all of these criteria are mutually exclusive: given a plausible set of premises, some can be derived from others.

There are other dimensions, apart from their distributive norms, in which theories of just entitlements differ. Some of these differences are logically implied by the nature of the norms themselves, while others are independent of them. One such dimension is the kinds of object to be distributed in conformity with a proposed criterion. Proffered items include all utility-producing goods, means of production, natural resources, the rents of superior skills or talents, and even human body parts. What one may do with the things to which one has strong entitlements – what weak entitlements and unprotected liberties one possesses – is largely a function of the sorts of thing to which others are strongly entitled. The intricate structure of permissibility, jointly formed by the rights one has against others and the rights others have against oneself, constitutes the fields of activity within which each person exercises those rights. It thereby also determines the respective spheres of market, state and charitable activities.

A third differentiating dimension is the range of subjects to be counted as having entitlements. Generally accepting the membership of all adult human beings in the class of title-holders, theories differ over whether their distributive norms extend to minors, members of other societies, deceased persons (in respect of bequest), persons conceived but not yet born (in respect of abortion), persons not yet conceived (in respect of capital accumulation and environmental

conservation) and non-human animals. Again, the nature and interpretation of a theory's distributive criterion often work to delimit its class of title-holders.

In the light of this multiplicity of differentiating dimensions, the classification – let alone assessment – of theories is no simple task. One, but by no means the only, important respect in which many of them can be compared is in terms of the scope they allow for unconstrained individual choice. Thus theories might be ranged along a spectrum from those that prescribe only an initial set of entitlements (permitting persons thereafter to dispose of these as they choose), to those that require constant enforceable adjustment of the content and distribution of entitlements to conform to certain norms. However, even this way of arraying competing theories is somewhat underspecified, inasmuch as it fails to capture the varied ramifications of the restrictions implied by different initial entitlements.

Hence it is an open question as to where on this spectrum one would locate theories that (via a unanimity requirement) construe each person's initial entitlement as a veto on a social or constitutional contract. Such an entitlement may in turn be derived from some interpretation of equal freedom, personal inviolability or interpersonal neutrality. Or it may itself be taken as an intuitively acceptable foundational postulate for deriving a more complex set of entitlements. Whether an initial contract theory is permissive or restrictive of wide individual choice depends upon its account of the terms of that contract. The derivation of these terms usually proceeds from some conception of human nature – of human knowledge and motivation – along with some meta-ethical theory about the nature of moral reasoning. Contractual terms generated by these premises may extend only to the design of political institutions, thereby leaving the determination of individuals' substantive entitlements to the legislative process. Alternatively, such contracts may stipulate a set of basic individual rights that are immune to legislative encroachment. In either case, the resultant scope for individual choice remains underdetermined. In the first case it depends upon the extent of legislation, while in the second it depends upon the size and nature of the stipulated set of rights. Laws and constitutional rights imply both restrictions on each person's conduct but also, *ipso facto*, restrictions on the extent of permissible interference with others' conduct.

Dispensing with the initial contract device and hypothetical unanimous agreement, some theories derive a set of entitlements directly (non-procedurally) from a substantive foundational value. Among such theories, one type assigns entitlements according to the differential incidence of some stipulated variable in the population of title-holders. Need and productivity are particularly prominent variables in this field, often acquiring their normative import from the values of welfare equalization and maximization. Clearly, applications of these distributive criteria respectively presuppose accounts of essential human requirements and of economic value. Although, for such theories, any shift in the incidence of the stipulated variable occasions a corresponding adjustment of entitlements, the issue of whether this adjustment must be imposed or occurs spontaneously partly turns on the model of interactive behaviour employed. In general, models indicating spontaneous adjustment generate that conclusion by ascribing dominance to altruistic (need) or income-maximizing (productivity) behaviour. To the extent that these ascription's are empirically unrealistic, such theories mandate enforceable restrictions on the scope for individual choice.

Another type of directly derived (non-contract-based) entitlement set is drawn from foundational values like equal freedom, personal inviolability or interpersonal neutrality, which, by definition, are of uniform non-differential incidence in the population of title-holders. Varying interpretations of these concepts tend nonetheless to converge on the Kantian injunction that persons must be treated as ends in themselves and, more specifically, that no person's ends may be systematically subordinated to those of another. Here the theoretical task is to design a set of entitlements that is independent of any particular conception of 'the good' – independent of particular preferences and (other) moral values – and that is such as to ensure that the consequences of persons' actions,

whether harmful or beneficial, are not imposed on others. A typical, though by no means invariable, structural feature of such an entitlement set is its extensive use of a threefold classification of things in the world as selves, raw natural resources and objects which are combinations of these. While title-holders are each vested with ownership of themselves (their bodies and labour), such theories often contain some sort of egalitarian constraint on individual entitlements to raw natural resources. The precise form of this constraint determines the nature of the encumbrances that may be imposed on the ownership of objects in the third category. But since these encumbrances exhaust the restrictions on what persons may do with what they own, such theories are presumed to allow considerable scope for individual choice.

It is hardly worth remarking that many theories of entitlement combine aspects of the three types outlined above. The assessment of competing theories – a complex task, as stated previously – commonly consists in testing for internal coherence and in appraising the interpretations placed on core concepts in the theory. Thus, if it is supposed that the moral principle underpinning a set of entitlements is that of justice, and that justice is analytically linked to the concept of rights, there is room for dispute as to whether the first (initial contract) and second (needs, productivity) types of theory are properly viewed as theories of entitlement. A distinctive normative feature of rights is that they are held non-contingently to confer an element of individuated discretion on their owners. It is unclear whether possession of a veto in a collective-choice procedure amounts to a sufficiently individuated sphere of discretion. On the other hand, the entitlements generated by considerations of need or productivity, while sufficiently individuated, appear to lack any necessarily discretionary character. A difficulty besetting the first and third types of theory arises with regard to the notion of initial entitlements. Specifically, it seems clear that the identification of each person's initial entitlement – either in a collective-choice procedure or under an egalitarian constraint on natural resource ownership – cannot be interpreted as an historically 'one-off' determination, in the face of an undecidable number and size of partially concurrent future generations. These are among the more salient problems commanding attention in current work on theories of entitlement.

See Also

- ▶ Economic Freedom
- ▶ Equality
- ▶ Inequality
- ▶ Justice
- ▶ Poverty
- ▶ Property Rights
- ▶ Redistribution of Income and Wealth

Bibliography

Buchanan, J.M. 1974. *The limits of liberty*. Chicago: University of Chicago Press.

Demsetz, H. 1964. Toward a theory of property rights. *American Economic Review, Papers and Proceedings* 57: 347–359.

Dworkin, R. 1981. What is equality? *Philosophy and Public Affairs* 10: 185–246; 283–345.

Hohfeld, W.N. 1919. *Fundamental legal conceptions*. New Haven: Yale University Press.

Lyons, D. (ed.). 1979. *Rights*. Belmont: Wadsworth Publishing Company.

Nozick, R. 1974. *Anarchy, state and utopia*. Oxford: Blackwell.

Rawls, J. 1971. *A theory of justice*, 1972. Oxford: Oxford University Press.

Sen, A.K. 1981. Rights and agency. *Philosophy and Public Affairs* 11: 3–39.

Steiner, H. 1987. *An essay on rights*. Oxford: Blackwell.

Entitlements in Laboratory Experiments

Sheryl Ball

Keywords

Altruism; Behavioural economics; Desert; Entitlement programmes; Entitlements; Entitlements in laboratory experiments; Fair allocation; Justice; Other-regarding behaviour;

Procedural fairness; Psychological games; Rawls, J.; Reciprocity; Self-interest

JEL Classifications
C9

Entitlements are rights granted by contract, law or practice. Under the assumption of pure self-interest, modelling games with entitlements is fairly straightforward; however, work in behavioural economics has consistently demonstrated the existence of other-regarding preferences, with strong effects of perceptions of what is fair. In the laboratory, behaviour is affected not only by the entitlement per se but also by the procedure by which entitlements come about. One form of laboratory entitlement is a more advantageous position in an economic game, where the advantage arises from a larger endowment, favourable exchange rules or greater decision-making authority. A second type of entitlement is a guaranteed payoff or a payoff floor. Experimental results show that the means by which entitlements are acquired is one cue that influences the nature of other-regarding behaviour. This is important both for understanding behaviour and the design of experiments.

In early experimental work on entitlements, Hoffman and Spitzer (1985) demonstrate that both the existence of an entitlement and its source determine economic outcomes. They study bilateral bargaining problems where one of the two subjects, called the 'controller', has unilateral authority to decide the outcome of a negotiation game in the event of disagreement. Authority is assigned based on either the outcome of a coin flip or the result of a simple test of a skill that is irrelevant to the experimental task. They find that controllers are most willing to exploit their power when they are assigned their role based on the skill test and are told that they 'earned' the right to be the controller – that is, that they have moral authority. These results are consistent with Burrows and Loomes (1994).

The subjects' behaviour illustrates Rawls's (1971) notion of 'desert', which requires that people deserve the conditions underlying their actions as well as the fruits of their actions. Thus subjects divided an endowment equally when the controller was chosen according to the flip of a coin and had low moral authority. On the other hand, both earning the right to be controller and higher moral authority triggered changes in observed allocations, so that outcomes favoured the controller. Entitlements that were earned or that involved 'morally unequal' agents were sufficient to trigger unequal outcomes. Equity theory developed by social psychologists is similar in spirit to this theory of justice.

Ideas of procedural fairness also affect perceptions of government entitlements. Fong (2001) looks at poll data on perceptions of poverty and opportunity, and finds that beliefs about others' effort, luck and opportunity play the largest role in determining support for government entitlement programmes. In particular these beliefs outweigh concerns about tax costs in supporting these programmes. These results are consistent with the experimental results discussed above, where low payoffs are acceptable if one displays low effort. If one's situation is determined by poor luck, however, one will give up some of one's earnings to increase the earnings of others.

A number of experimental studies on income redistribution examine Rawls's claim that individuals prefer an income redistribution rule that maximizes the position of the poorest member of society (Frohlich and Oppenheimer 1990). Studies where subjects must choose a principle of distributive justice and a tax system in addition to participating in a production task find that people choose rules that maximize the productivity of society while maintaining a minimum floor for the worst off members. Subjects generate greater output in experiments where they are able to determine the entitlements for the worst off individual in their group, again demonstrating that the source of entitlements matters.

These results show that researchers need to pay attention to how entitlements are determined. This is a complication for theories of behavioural economics or psychological games. People do not have a pure taste for fair allocations; they are more self-interested, altruistic or fair according to circumstances that depend on how advantage arises. This behaviour is closely related to

reciprocity, but that is often modelled as 'if you are nice to me I'll be nice to you' (Bowles and Gintis 2001). In contrast, this collection of results can be interpreted as, 'I will respect your entitlement if you deserve it'.

A preference for procedural factors also complicates experimental design, since subjects behave in a more self-interested manner when entitlements are earned than when they are randomly assigned. Researchers must be careful to consider how subjects will interpret the rules by which advantages are assigned or they may risk introducing nuisance variables. Future work might deliberately award entitlements in a manner that subjects view as unjust to see whether that produces yet another pattern of behaviour.

See Also

- ▶ Behavioural Game Theory
- ▶ Coase Theorem
- ▶ Experimental Economics
- ▶ Fair Allocation
- ▶ Justice
- ▶ Psychological Games

Bibliography

Burrows, P., and G. Loomes. 1994. The impact of fairness on bargaining behavior. *Empirical Economics* 19: 201–221.

Bowles, S. and Gintis, H. 2001. The inheritance of economic status: Education, class and genetics. In *Genetics, behavior and society,* ed. M. Feldman. In *International encyclopedia of the social and behavioral sciences,* ed. N. Smelser and P. Baltes. Oxford: Elsevier.

Fong, C. 2001. Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics* 82: 225–246.

Frohlich, N., and J.A. Oppenheimer. 1990. Choosing justice in experimental democracies with production. *American Political Science Review* 84: 461–477.

Hoffman, E., and M.L. Spitzer. 1985. Entitlements, rights and fairness: An experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies* 14: 259–297.

Lissowski, G., T. Tyszka, and W. Okrasa. 1991. Principles of distributive justice: Experiments in Poland and America. *Journal of Conflict Resolution* 35: 98–119.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Belknap, Harvard University Press.

Entrepreneur

Mark Casson

There are several theories of the entrepreneur, but very few mathematical models which formally analyse entrepreneurial behaviour within a closed economic system. Indeed, it is often argued that by its very nature entrepreneurial behaviour cannot be predicted using deterministic models. Entrepreneurship, it is claimed, is essentially a spontaneous and evolutionary phenomenon.

The term 'entrepreneur' seems to have been introduced into economic theory by Cantillon (1755), but the entrepreneur was first accorded prominence by Say (1803). It was variously translated into English as 'merchant', 'adventurer' or 'employer', though the precise meaning is the undertaker of a project. John Stuart Mill (1848) popularized the term in England, though by the turn of the century it had almost disappeared from the theoretical literature (though see Marshall 1890).

The 'disappearance' of the entrepreneur is associated with the rise of the neoclassical school of economics. The entrepreneur fills the gap labelled 'fixed factor' in the neoclassical theory of the firm. Entrepreneurial ability is analogous to a fixed factor endowment because it sets a limit to the efficient size of the firm. The static and passive role of the entrepreneur in the neoclassical theory reflects the theory's emphasis on perfect information – which trivializes management and decision-making – and on perfect markets – which do all the coordination that is necessary and leave nothing for the entrepreneur (cf. Baumol 1968).

According to Schumpeter (1934), the entrepreneur is the prime mover in economic development, and his function is to innovate, or 'carry out new combinations'. Five types of innovation are distinguished: the introduction of a new good (or an improvement in the quality of an existing good), the introduction of a new method of production, the opening of a new market – in particular an export market in new territory – the

'conquest of a new source of supply of raw materials or half-manufactured goods' and the creation of a new type of industrial organization – in particular the formation of a trust or some other type of monopoly. Schumpeter is also very clear about what the entrepreneur is *not*: he is not an inventor, but someone who decides to allocate resources to the exploitation of an invention; nor is he a risk-bearer: risk-bearing is the function of the capitalist who lends funds to the entrepreneur. Essentially, therefore, Schumpeter's entrepreneur has a managerial, or decision-making role.

This view receives qualified support from Hayek (1937) and Kirzner (1973), who emphasize the role of the entrepreneur in acquiring and using information. The entrepreneur's alertness to profit-opportunities, and his readiness to exploit them through arbitrage-type operations, makes him the key element in the 'market process'. Hayek and Kirzner regard the entrepreneur as responding to change – as reflected in the information he receives – whilst Schumpeter emphasized the role of the entrepreneur as a source of change. These two views are not incompatible though: a change effected by one entrepreneur may cause spill-over effects which alter the environment of other entrepreneurs. Hayek and Kirzner do not insist on the novelty of entrepreneurial activity, however, and it is certainly true that a correct decision is not always a decision to innovate; premature innovation may be commercially disastrous. Schumpeter begs the question of whether someone who is the first to evaluate an innovation, but decides (correctly) not to innovate, qualifies as an entrepreneur.

Leibenstein (1968) regards the entrepreneur as someone who achieves success by avoiding the inefficiencies to which other people – or the organization to which they belong – are prone. The main virtue of Leibenstein's approach is that it emphasizes that, in the real world, success is exceptional and failure is the norm.

Knight (1921) insists that decision-making involves uncertainty. Each business situation is unique, and the relative frequencies of past events cannot be used to evaluate the probabilities of future outcomes. According to Knight, measurable risks can be diversified – or 'laid off' – through insurance markets, but uncertainties cannot. Those who take decisions in highly uncertain environments must bear the full consequences of those decisions themselves. These people are entrepreneurs: they are the owners of businesses and not the salaried managers that make day-to-day decisions.

It is not clear, at first sight, whether there is any common thread which runs through these various theories of the entrepreneur. Casson (1982) attempts to identify a shared element by introducing the concept entrepreneurial judgement. The entrepreneur is defined as someone who specializes in taking judgemental decisions about the allocation of scarce resources. The essence of a judgemental decision is that there is no decision rule that can be applied that is both obviously correct and involves using only freely available information. Suppose, for example, that a decision rule, is used; then there must be some initial judgement that the chosen rule, and not some other rule, is the appropriate one. No rule can ever be fully self-justifying: there is no definitive model which demonstrates that one rule is always superior to another. Ultimately, the justification for a rule must be some property of the environment, which in many cases cannot be observed.

It is evident that this concept of judgemental decision-making rejects the 'naive neoclassical' view that all decision-making merely involves marginalist calculations based upon public information supplied by the price system. It recognizes that not only is information costly, but that the costs of acquiring information are different for different people. Furthermore, because their access to information differs, different people will make different decisions in the same situation. The essence of judgemental decision-making is that the outcome depends upon *who* makes the decision.

When judgements differ, confident individuals can back their own judgement by taking up speculative positions against other people who hold a conventional view. The confident individuals 'bet' against others by acquiring assets that they believe other people have under-valued, disposing of assets that they believe other people have over-valued, undertaking projects that other people do

not consider profitable, and so on. Using this approach, the arbitraging activity described by Hayek and Kirzner, and the innovative activity described by Schumpeter, are seen to be special cases of the general concept of entrepreneurial speculation based upon self-confident judgement.

In a market economy, individuals who lack confidence in their judgement can delegate decisions to entrepreneurs. The individual entrusts his wealth to an entrepreneur, who allocates this wealth in accordance with his own judgement. In practice, the individual will often diversify his risks by using a 'portfolio' of different entrepreneurs. The delegation of decision-making can be effected in various ways. An individual may supply capital at fixed interest to an entrepreneur who is self-employed or is the owner-manager of a firm. He may own an equity stake in a firm where the entrepreneur acts partly as a salaried employee; or he may deposit his funds in a bank whose managers advance loans to firms and self-employed entrepreneurs.

To overcome the principal-agent problems involved in the delegation of decisions, it is normally necessary for the supplier of finance not only to have confidence in the entrepreneur's judgement, but also to trust the entrepreneur to exercise this judgement in pursuit of maximum profit. Unless the entrepreneur has an established reputation, he has a strategic problem in obtaining the confidence of others. Because of the differences in judgement mentioned earlier, the entrepreneur will normally be more optimistic about a project than are his potential financial backers. His backers will therefore perceive higher risks, and set a higher cost of capital than the entrepreneur believes is warranted. If, however, he persuades his backers to share his optimism, then they may preempt his project, since they already have the finance to proceed with it and he does not.

This leads directly to the question of trust. Just as the financiers must trust the entrepreneur to use his funds in their interests, so the entrepreneur must trust his financial backers not to preempt his project for themselves. Part of the problem can be solved by using an 'honest broker' such as a bank, which vets entrepreneurial projects on behalf of investors but ties its own hands by not entering into entrepreneurial projects on its own account. In countries where the banking system is underdeveloped, the extended family often fulfils a similar function of 'honest broking' between the older generation who are potential investors and the younger generation who are potential entrepreneurs. Another method of building trust is to supply finance in a sequence of small instalments so that both parties have an incentive to behave honourably in order not to put future relations between them at risk.

Much of the information required for decision-making is not merely costly to obtain, but is not available by direct observation at all. Another way of saying this is that decisions are governed not only by objective information but also by subjective beliefs. An individual's beliefs originate with his culture and his religion as well as with his direct experience of life. Some cultures appear to give greater encouragement to entrepreneurship than others. A culture which stresses individuality rather than conformity encourages an individual to form an independent judgement of a situation. A culture which emphasizes human autonomy rather than fatalistic submission to nature encourages the kind of self-confidence required of the entrepreneur. A culture which emphasizes the heroic aspects of leadership rather than the corrupting effects of power-seeking encourages individuals to undertake ambitious projects which call for a high degree of organization, and so on. Cultural values have always been emphasized in the literature on entrepreneurship: Schumpeter, for example, refers to the dream and will to found a private dynasty, the will to conquer and the joy of creating, while Weber (1930) emphasizes the Protestant ethic and the concept of calling and Redlich (1956) the militaristic values of the 'captains of industry'. Writers on business history almost invariably stress the influence of culture and personality on the behaviour of the entrepreneur.

A common criticism of theories which place considerable weight on cultural characteristics and personality traits is that they are difficult to test. Indeed, it is often suggested that because the behaviour of individual entrepreneurs tends to be unpredictable, theories of entrepreneurship are

untestable. It is, however, quite possible that while the behaviour of individual entrepreneurs cannot be predicted, the behaviour of entrepreneurs as a group is predictable. Furthermore, a theory of entrepreneurship may generate propositions relating to other social phenomena besides the behaviour of entrepreneurs themselves. With certain qualifications, it is possible to develop a model in which both the level of entrepreneurial activity and the functional distribution of income between entrepreneurship and other factors are simultaneously determined.

Given that the entrepreneur specializes in judgemental decision-making, it is possible to formulate a derived demand for entrepreneurial services which varies according to the demands which the business environment places upon judgement. The more complex the environment, the faster the pace of change, and the more radical the structural adjustments that these changes call for, the greater will be the demand for entrepreneurs. A large demand for entrepreneurs will be reflected in substantial profit opportunities for people who can anticipate changes and correctly foresee their consequences. Individual profit opportunities will not be competed away so long as each opportunity can be preempted by a single entrepreneur before others come to form the same judgement as he has done. In the short run, therefore, the successful entrepreneur can earn a monopolistic rent to superior judgement.

In the long run, however, entry into entrepreneurship will tend to compete away any expected return to entrepreneurial activity which exceeds the expected return to non-entrepreneurial activity (after due allowance for different levels of risk and for the non-pecuniary net benefits of the two kinds of activity). The competing away of entrepreneurial rents may not be complete, however, because access to capital may prove a barrier to entry for the reasons explained above.

Long-run entry corresponds to a movement along a long-run supply curve for entrepreneurs. The total supply of entrepreneurs is measured by the number of individuals whose principal activity is to exercise their judgement to allocate resources. The people concerned may be senior salaried managers or the self-employed – given the definition above, it is impossible to identify the entrepreneur simply by his contractual status in employment. An increase in the supply of entrepreneurs is effected by individuals transferring out of manual work and non-entrepreneurial decision-making (i.e. routine management), and from unemployment and leisure, and by net inward migration of entrepreneurs from abroad. The position and elasticity of the entrepreneurial supply curve depends upon the expected return to non-entrepreneurial activity abroad, the distribution of judgemental ability within the indigenous population, cultural attitudes, and barriers to entry and exit which reduce mobility between the entrepreneurial and non-entrepreneurial groups.

Given both the long run supply and long run demand for entrepreneurs, it is possible to visualize a long equilibrium in which the marginal entrepreneur earns an approximately normal return, intra-marginal entrepreneurs earn a quasi-rent to superior judgement, and intra-marginal non-entrepreneurs earn quasi-rents for their non-entrepreneurial abilities. The equilibrium return to entrepreneurship, and the equilibrium number of entrepreneurs, depend upon the parameters of the demand and supply curves, as described above. This equilibrium is a partial equilibrium, conditional upon the returns to non-entrepreneurial activity within the economy. It is also possible to derive a general equilibrium by endogenizing the return to non-entrepreneurial activity.

It should be emphasized, however, that any kind of ‘equilibrium’ in a ‘market for entrepreneurs’ is essentially an analytical fiction because the adjustment of this market to an equilibrium is itself an entrepreneurial task. The decision whether to hire an entrepreneur, and the decision whether to become one, are both entrepreneurial decisions. It is difficult for entrepreneurs to intermediate in the market for entrepreneurs because it is difficult to buy and sell ‘human capital’ of this kind. To introduce a Walrasian auctioneer to coordinate supply and demand decisions in the market for entrepreneurs would be self-contradictory, for it is only because of the absence of the Walrasian auctioneer that entrepreneurs are required in the first place. Thus while the concept of a market

equilibrium for entrepreneurs is a useful analytical device, it is erroneous to suppose that the market for entrepreneurs is ever in a full equilibrium.

Twenty years ago, the study of the entrepreneur was regarded as a 'gap' in economic theory. It is now recognized that this gap cannot be filled without radically changing the nature of the theory itself. The entrepreneur can only be understood properly in the context of an economic model which does full justice to the structural complexity and the evolutionary nature of the economy (Nelson and Winter 1982). Within such a model the 'equilibrium' concept remains a useful analytical device, but one of limited practical relevance. The study of the entrepreneur leads to a vision of economics much wider than that of a subject which parsimoniously derives a consistent set of price and quantity equations. Aspects of human personality – such as self-confidence – acquire a crucial role – and so too does the malleability of the personality under the influence of cultural attitudes. The theory of the entrepreneur, therefore, is not the last step which renders the conventional theory of value complete, but the first step towards an economic theory which forms part of a wider integrated body of social science.

See Also

- ▶ Codetermination and Profit-Sharing
- ▶ Corporate Economy
- ▶ Interest and Profit
- ▶ Profit and Profit Theory

Bibliography

Baumol, W.J. 1968. Entrepreneurship in economic theory. *American Economic Review: Papers and Proceedings* 58: 64–71.

Cantillon, R. 1755. *Essai sur la nature du commerce en général,* ed. H. Higgs. London: Macmillan, 1931.

Casson, M.C. 1982. *The entrepreneur: An economic theory*. Oxford: Martin Robertson.

Hayek, F.A. von. 1937. Economics and knowledge. *Economica*, NS 4: 33–54.

Kirzner, I.M. 1973. *Competition and entrepreneurship*. Chicago: University of Chicago Press.

Knight, F.H. 1921. *Risk, uncertainty and profit,* ed. G.J. Stigler. Chicago: University of Chicago Press, 1971.

Leibenstein, H. 1968. Entrepreneurship and development. *American Economic Review* 58: 72–83.

Marshall, A. 1890. *Principles of economics*, 9th edn, 2 vols, ed. G.W. Guillebaud. London: Macmillan, 1961.

Mill, J.S. 1848. *Principles of political economy*, New edn, ed. W.J. Ashley. London: Longmans, 1909.

Nelson, R.R., and S.G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.

Redlich, F. 1956. The military enterpriser: A neglected area of research. *Explorations in Entrepreneurial History*, Series 1, 8: 252–256.

Say, J.B. 1803. *A treatise on political economy: or the production, distribution and consumption of wealth*. New York: Augustus M. Kelley, 1964.

Schumpeter, J.A. 1934. *The theory of economic development*. Trans. R. Opie. Cambridge, MA: Harvard University Press.

Weber, M. 1930. *The protestant ethic and the spirit of capitalism*. Trans. T. Parsons. London: George Allen & Unwin.

Entrepreneurship

William J. Baumol and Melissa A. Schilling

Abstract

This article describes the recent expansion of research on entrepreneurship, innovation and growth. Although the entrepreneur is widely credited with critical contributions to innovation and growth, the subject of entrepreneurship has virtually disappeared from mainstream theory and standard textbooks. Reasons explaining this gap are indicated. In addition to some brief materials on earlier writings, the rich body of recent work on the subject, both theoretical and empirical, is surveyed, illustrating the wide variety of subjects explored and the insights offered by the new literature.

Keywords

Baumol, W.; Cantillon, R.; economic growth; entrepreneurship; equilibrium; innovation; Kirzner, I.; Knight, F.; Marshall, A.; Mill, J.S.; Neoclassical theory; Opportunity costs;

Patents; Product development; Property rights; Risk; Say, J.-B.; Schumpeter, J.; Self-employment; Social status; Transfer of technology; Uncertainty premium; Venture capital

JEL Classifications
L2; O12; O31; O33

An entrepreneur is an individual who organizes, operates, and assumes the risk of creating new businesses. There are two types. A replicative entrepreneur organizes a new business firm that is like other firms already in existence. An innovating entrepreneur provides something new – a new product or process, or a new type of business structure, a new approach to marketing, and so on. These innovations need not be productive or beneficial. For example, Richard Cantillon (one of the first great economic theorists) spoke of thieves who are entrepreneurs (Cantillon 1730, pp. 54–5). And Joseph A. Schumpeter, arguably the contributor of the most important analysis of entrepreneurship, included as an entrepreneurial act '.. .the creation of a monopoly position (for example, through trustification).. ..' (Schumpeter 1911, p. 66). Entrepreneurs (interpreted as the self-employed) are estimated to constitute about seven percent of the labour force in the United States (U.S. Bureau of the Census 2004). Most of them are probably replicative, not innovative, entrepreneurs.

It is widely agreed that the entrepreneur plays an important role in economic growth. But the evidence shows little correlation between an economy's number of replicative entrepreneurs and its growth rate. Innovative entrepreneurs do make a substantial difference to a nation's growth rate, having introduced many breakthrough innovations like the telephone and the airplane. The primary social contribution of replicative entrepreneurship is as a means for individuals to escape poverty, because such undertakings require little capital, education or experience. Still, the data show that entrepreneurs, on average, earn less than employees with similar education and experience (Freeman 1978; Astebro 2003; Benz and Frey 2004).

Although economists have recently exhibited a resurgence of interest in entrepreneurship, the entrepreneur nevertheless rarely shows up in contemporary mainstream economic theory.

Early Writings and the Origin of the Term

Until the 20th century, writings in English referred to entrepreneurs as 'adventurers' or 'undertakers' (see, for example, Marshall 1923, p. 172). Apparently, the term 'entrepreneur' was introduced by Cantillon in the French translation of his great work, *Essai Sur la Nature de Commerce en Général* (1730, p. 54), but what is apparently *his* English text uses the word 'undertaker'. The early writings on entrepreneurship were descriptive rather than theoretical. Cantillon's discussion (1730, ch. 11) is brief, focusing on replicative entrepreneurs: '.. .wholesalers in Wool and Corn, Bakers, Butchers, Manufacturers and Merchants of all kinds.. ..' (1730, p. 51). Cantillon's main point, like that of Frank H. Knight (1921), was the task's riskiness: 'These Undertakers can never know how great will be the demand in their City, nor how long their customers will buy of them since their rivals will try all sorts of means to attract customers from them. All this causes so much uncertainty among these Undertakers that every day one sees some of them become bankrupt' (1730, p. 51).

Nearly a century later, Jean-Baptiste Say's (1819) discussion is still brief, but richer. Say seems interested primarily in innovating entrepreneurs, dealing with three types of 'producers': scientists, entrepreneurs and labourers. Using mechanical locks as an example, the scientist investigates '.. .the properties of iron, the method of extracting from the mine and refining the ore.. .' The entrepreneurs deal with '.. .application of this knowledge to a useful purpose.. .,' while the third group – the workers – actually make the product (1819, p. 80). And any successful economy needs all three: 'Nor can [industry] approximate to perfection in any nation, till that nation excel in all three branches' (1819, p. 80).

Thus, Say blames poverty in Africa on the absence of scientists and entrepreneurs. Lack of

entrepreneurs alone can undercut prosperity, even with scientific knowledge abundant, for without the entrepreneur, '.. .that knowledge might possibly have lain dormant in the memory of one or two persons, or in the pages of literature' (1819, p. 81). This is precisely the explanation that one of the present authors proposed for the failure of medieval China and the Soviet Union to translate an abundance of non-military inventions into viable consumer products (Baumol 2002, chs. 5, 14). Say also foreshadows some of Schumpeter's analysis (see below): 'In manufacture.. .if success [in innovation] ensue, the adventurer is rewarded by a longer period of exclusive advantage, because his process is less open to observation' (1807, p. 84).

Finally, Say mentions the spillovers of innovation and their justification for governmental financing: 'The charges of experiment, when defrayed by the government... [are] hardly felt at all, because the burthen is divided among innumerable contributors; and the advantages resulting from success being a common benefit to all, it is by no means inequitable that the sacrifices, by which they are obtained, should fall on the community at large' (1819, p. 85).

Before Schumpeter's breakthrough (see below), the subject was touched upon by economists like J.S. Mill, Alfred Marshall and (a bit later) Knight. Generally, their focus was not on innovative entrepreneurship, and they emphasized management's directing of going concerns rather than establishment of new firms. (But Marshall 1923, p. 172, does digress briefly to mention Matthew Boulton's significant role as an entrepreneur dealing with James Watt's inventions.) Today, however, these discussions would hardly be considered theory. Rather, they are usually narratives containing illuminating observations. They assert that the entrepreneur's payment is a residual after other inputs are compensated, and that compensation is determined by the entrepreneur's ability and the supply of entrepreneurship in the market. They note that entrepreneurs employ themselves, so that unlike other inputs there is no demand function, as for other inputs.

Disappearance of the Entrepreneur from Modern Mainstream Economics

Given the acknowledged importance of the entrepreneur's role, it could be hoped that modern theoretical economics, with its powerful analytic tools, would have produced an extensive entrepreneurship analysis. Instead, the opposite happened – the entrepreneur became the 'invisible man' in mainstream theory. There are at least two reasons for this. First, the most advanced and powerful microeconomic models predominately study timeless static equilibria. But, for the entrepreneur, the transition process is the heart of the story. Schumpeter (1911) shows the entrepreneur as a destroyer of equilibria by constant innovation, while Israel Kirzner (1979) tells how the alert entrepreneur seeks out the arbitrage opportunities presented by disequilibria, thereby moving the economy back toward equilibrium. Such a relentless attack upon both equilibria and disequilibria does not fit a stationary model from which firm creation and invention are excluded.

The second reason for the entrepreneur's disappearance from mainstream theory is that, by definition, an invention is something never available before. So invention is the ultimate heterogeneous product. This impedes the optimality analysis underlying most microeconomic theory. Explicitly or implicitly, an optimality calculation entails a comparison among possible substitute choices, while the innovating entrepreneur normally deals with no well-defined substitutes with quantifiable attributes. In contrast, the standard theory of the firm analyses repetitious decisions of management in fully operational enterprises where the entrepreneur has already completed his job and left to create other firms.

Thus, neoclassical theory is justified in excluding the entrepreneur, because it deals with subjects for which the entrepreneur is irrelevant. That does not mean that no theory of entrepreneurship is needed, or that such a theory is lacking, but it means that a theory of entrepreneurship must be sought elsewhere, and that is what Schumpeter succeeded in doing.

Brief Summary: Schumpeter's Model: The Supply and Earnings of Entrepreneurial Activity

The basic Schumpeterian model (1911) notes that the successful innovative entrepreneur's reward is profit temporarily exceeding that of perfect competition. This attracts rivals who seek to share those profits by imitating the innovation, and thereby erode its super-competitive earnings. To prevent termination of these rewards, the entrepreneur can never desist from further innovation and cannot rest on his laurels.

Perhaps most important, the Schumpeterian analysis shows how the entrepreneur is driven to work without let-up for economic growth. Thus, it clearly reveals the tight association between innovative entrepreneurship and growth.

Allocation Between Productive and Unproductive Entrepreneurship

Some work of one of the present authors (Baumol 2002, ch. 14) tells much of the rest of the story about the supply and allocation of productive entrepreneurship and the key role of evolving institutions. In the economic growth literature, it has often been asserted that an expanded supply of entrepreneurs effectively stimulates growth, while shrinkage in the supply undermines growth. But the standard explanation of the entrepreneurs' appearance and disappearance is shrouded in mystery, with hints about cultural developments and vague psychological and sociological changes. The historical evidence suggests a more mundane explanation: that entrepreneurs are always present but, as the structure of rewards in the economy changes, entrepreneurs switch their activities, moving to where payoffs become more attractive. In doing so, they move in and out of the activities usually recognized as entrepreneurial, exchanging them for other activities that also require enterprising talent but are often distant from production of goods and services. The generals of ancient Rome, the Mandarins of the Tang, Sung, and Ming Chinese empires, the captains of late medieval private and mercenary armies, the rent-seeking contemporary lawyers, and the Mafia Dons – all are clearly enterprising and often successful. And when institutions have changed so as to modify profoundly the relative payoffs offered by the different enterprising activities, the supply of entrepreneurs has shifted accordingly. Here, it is helpful to distinguish two categories of entrepreneurs, the productive and the unproductive entrepreneurs, with the latter, in turn, divided into subgroups such as rent-seeking entrepreneurs and destructive entrepreneurs, including the organizers of private armies or criminal groups. Once there is a pertinent change in the institutions that govern the relative rewards, the entrepreneurs will shift their activities between productive and unproductive occupations, so the set of productive entrepreneurs will appear to expand or contract autonomously. For example, when institutions change to prohibit private armies, entrepreneurs are led to look elsewhere to realize their financial ambitions. If, simultaneously, rules against confiscation of private property and for patent protection of inventions are adopted, entrepreneurial talent will shift into productive, innovative directions.

Recent Studies: Other Disciplines and Empirical Approaches

Outside mainstream economic analysis, research on entrepreneurship has expanded rapidly since the 1980s, particularly that by specialists in management, psychology, and sociology. We focus here on three streams of work that have attracted the most scholarly attention: (*a*) how differences among individuals influence entry into (and success in) entrepreneurship, (*b*) how environment influences entrepreneurship, and (*c*) the strategies and forms of organization used by entrepreneurs.

Differences Among Individuals

There are numerous studies investigating how differences among individuals (in attributes such

as education, age, experience, social position and psychology) are associated with a propensity to become self-employed and the likelihood of success at entrepreneurship. A wide variety of studies have indicated that individuals with higher education than the general population are more likely to become entrepreneurs (Shane 2003). Robinson and Sexton (1994) and others have found that number of years of education is significantly related to likelihood of becoming self-employed, and Bates (1995) found that individuals with a graduate education were significantly more likely to become self-employed. Age appears to have an inverted U-shaped relationship with likelihood of forming a new venture. Entrepreneurship first increases with age because of experience, and then decreases with age because of opportunity costs and uncertainty premiums (Bates 1995; Shane 2003).

A number of studies that look at how experience influences likelihood of starting a business and the success of the new venture have found that general business experience (Evans and Leighton 1989; Robinson and Sexton 1994), experience specific to the industry in which the entrepreneur later founds a business (Aldrich 1999), and prior self-employment (Carroll and Mosakowski 1987) all increase the likelihood that an individual will found a new business. Furthermore, such experience tends to improve new venture performance and survival rates (Gimeno et al. 1997).

Studies have revealed that, in general, social status increases the likelihood of forming a new venture (for example, Stuart et al. 1999). The number and diversity of an individual's social ties also increase the likelihood of founding a company (Aldrich et al. 1987), as well as the success of the venture (Hansen 1995). Psychological factors also influence an individual's likelihood of becoming an entrepreneur (Shane 2003). In particular, extraversion (Babb and Babb 1992), need for achievement (Hornaday and Aboud 1973), risk-taking propensity (Astebro 2003), self-efficacy (Zietsma 1999), overconfidence (Arabsheibani et al. 2000), and creativity (Ames and Runco 2005) have all been shown to be significantly related to an individual's likelihood of becoming an entrepreneur.

Environmental Factors

A number of industry characteristics influence new venture formation. Market size (Pennings 1982) and growth (Dean and Meyer 1992) increase the likelihood of new firm formation, while uncertainty from technological change decreases the rate of business start-ups (Audretsch and Acs 1994). Capital intensity also reduces new firm formation by raising entry costs (Dean and Meyer 1992). The density of firms has an inverted U-shaped relationship with new firm formation (Carroll and Wade 1991). Too few firms in an industry may signal that there is no opportunity worth pursuing, or scarcity of market information. Thus, initial increases in the density of firms in the industry encourage business start-ups (Shane 2003), although high density can increase competition for resources and create an entry barrier.

Not surprisingly, the institutional environment of an industry or region also affects new firm formation. Capital availability (for example, low-cost debt or venture capital) enhances firm formation (McMillan and Woodruff 2002). Higher marginal federal income tax rates decrease self-employment (Gentry and Hubbard 2000) and business tax concessions increase business start-ups (Dana 1987).

Stronger property rights encourage entrepreneurship, presumably because they assure entrepreneurs that they can appropriate the fruits of their efforts (McMillan and Woodruff 2002). Researchers have also investigated the role of university technology-transfer offices on entrepreneurship, with most research indicating that such offices increase rates of new venture formation, particularly when technology-transfer offices are structured to profit from the transfers (Markman et al. 2005). Finally, socio-cultural norms about the desirability of self-employment or the risks of failure are significantly related to rates of business start-ups in a nation or ethnic group (Butler and Herring 1991).

Strategy and Organization

The area of entrepreneurial strategy that has received most research attention is method of financing. Consistent with Knight's (1921) argument that self-financing is needed to overcome

moral hazard problems, most entrepreneurs finance their ventures primarily with their own capital (Aldrich 1999; Shane 2003). However, funds provided by 'angel' investors (wealthy individuals who invest in entrepreneurial companies, usually at an early stage) and venture capitalists are also important. The research on angel investment is sparse, but there is more research on venture capitalst investment. A number of researchers have investigated how venture capitalists choose their investments, mitigate risk, and influence new venture survival and growth (Bygrave and Timmons 1992). Some studies have also examined how entrepreneurs identify opportunities (Shane 2003), their degree of reliance on patent protection (Shane 2001), the effect of entrepreneurs' new product development strategies (Zahra and Bogner 2000), and their breadth of market focus (Bhide 2000; Gimeno et al. 1997).

Finally, there also has been some research on the organization of new ventures how they are formed as legal entities, the performance implication of this choice (Delmar and Shane 2004), and the effect of venture team size and background (Eisenhardt and Schoonhoven 1990). In general, formation as a legal entity and a large, diverse venture team appear to improve new venture performance.

On the State of the Theory of Entrepreneurship

Our discussion demonstrates that the beginnings of a significant theory of entrepreneurship already exist. The analysis uses little mathematics to derive any formal theorems, and its results are primarily qualitative. But this nascent theory of entrepreneurship does tell us about its supply and earnings, its role in the pricing of its products and the role of the price mechanism in its allocation among alternative activities. The Schumpeterian model tells us about the determination of entrepreneurs' profits and the prices of their products, as well as their influence on the supply of their activity. The model of productive and unproductive entrepreneurship tells us more about supply, as well as about the allocation of this resource. The empirical research adds further insight into the factors that increase the likelihood of individuals engaging in, and being successful at, entrepreneurship.

Beyond the stationary analysis of standard microeconomic theory, we see that the entrepreneurship models enable us to deal with such important questions as what features of the structure of the free market economy have caused it to outperform by an order of magnitude the innovation and growth of any alternative economic system. The institutional changes that reallocated much of entrepreneurship from redistributive to productive activities are, according to the model, the key to the answer. And this has profound policy implications both for developing countries seeking desperately to escape their poverty and for developed economies seeking to keep up the pace of their growth.

See Also

- ▶ Cantillon, Richard (1697–1734)
- ▶ Growth and Institutions
- ▶ Intellectual Property
- ▶ Knight, Frank Hyneman (1885–1962)
- ▶ Schumpeterian Growth and Growth Policy Design

Bibliography

Aldrich, H. 1999. *Organizations Evolving*. London: Sage.

Aldrich, H., B. Rosen, and W. Woodward. 1987. The impact of social networks on business foundings and profit: A longitudinal study. In *Frontiers of entrepreneurship research*, ed. N. Churchill et al. Babson Park: Babson College.

Ames, M., and M. Runco. 2005. Predicting entrepreneurship from ideation and divergent thinking. *Creativity and Innovation Management* 14: 311–315.

Arabsheibani, G., D. De Meza, J. Maloney, and B. Pearson. 2000. And a vision appeared unto them of a great profit: Evidence of self-deception among the self-employed. *Economics Letters* 67: 35–41.

Astebro, T. 2003. The return to independent invention: Evidence of unrealistic optimism, risk seeking or skewness loving. *Economic Journal* 113: 226–238.

Audretsch, D., and Z. Acs. 1994. New firm startups, technology, and macroeconomic fluctuations. *Small Business Economics* 6: 439–449.

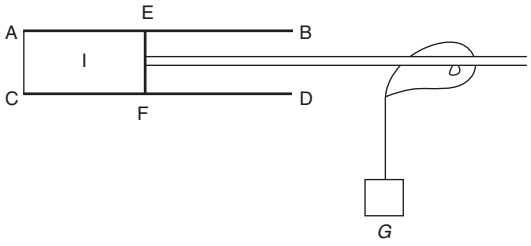
Babb, E., and S. Babb. 1992. Psychological traits of rural entrepreneurs. *Journal of Socio-Economics* 21: 353–362.

Bates, T. 1995. Self-employment entry across industry groups. *Journal of Business Venturing* 10: 143–156.
Baumol, W. 2002. *The free-market innovation machine: Analyzing the growth miracle of capitalism*. Princeton: Princeton University Press.
Benz, M., and B. Frey. 2004. Being independent raises happiness at work. *Swedish Economic Policy Review* 11: 95–134.
Bhide, A. 2000. *The origin and evolution of new businesses*. New York: Oxford University Press.
Butler, J., and C. Herring. 1991. Ethnicity and entrepreneurship in America: Toward an explanation of racial and ethnic group variations in self-employment. *Sociological Perspectives* 34: 79–95.
Bygrave, W., and J. Timmons. 1992. *Venture capital at the crossroads*. Boston: Harvard Business School Press.
Cantillon, R. 1730. *Essai Sur la Nature de Commerce en Général*. Trans. H. Higgs, 1931 London: Macmillan.
Carroll, G., and E. Mosakowski. 1987. The career dynamics of self employment. *Administrative Science Quarterly* 32: 570–589.
Carroll, G., and J. Wade. 1991. Density dependence in organizational evolution of the American brewing industry across different levels of analysis. *Social Science Research* 20: 271–302.
Dana, L. 1987. Entrepreneurship and value creation – an international comparison of five commonwealth nations. In *Frontiers of entrepreneurship research*, ed. N. Churchill et al. Babson Park: Babson College.
Dean, T., and G. Meyer. 1992. New venture formation in manufacturing industries: A conceptual and empirical analysis. In *Frontiers of entrepreneurship research*, ed. N. Churchill et al. Babson Park: Babson College.
Delmar, F., and S. Shane. 2004. Legitimating first: Organizing activities and the survival of new ventures. *Journal of Business Venturing* 19: 385–410.
Eisenhardt, K., and K. Schoonhoven. 1990. Organizational growth: Linking founding team, strategy, environment, and growth among U.S. semiconductor ventures, 1978–1988. *Administrative Science Quarterly* 35: 504–529.
Evans, D., and L. Leighton. 1989. Some empirical aspects of entrepreneurship. *American Economic Review* 79: 519–535.
Freeman, R. 1978. Job satisfaction as an economic variable. *American Economic Review* 68: 135–141.
Gentry, W., and R. Hubbard. 2000. Tax policy and entrepreneurial entry. *American Economic Review Papers and Proceedings* 90: 283–292.
Gimeno, J., T. Folta, A. Cooper, and C. Woo. 1997. Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms. *Administrative Science Quarterly* 42: 750–783.
Hansen, E. 1995. Entrepreneurial networks and new organization growth. *Entrepreneurship Theory and Practice* 19(4): 7–19.
Hornaday, J., and J. Aboud. 1973. Characteristics of successful entrepreneurs. *Personnel Psychology* 24: 141–153.
Kirzner, I. 1979. *Perception, opportunity and profit*. Chicago: University of Chicago Press.
Knight, F. 1921. *Risk, uncertainty and profit*. Boston/New York: Houghton Mifflin Company.
Markman, G., P. Phan, D. Balkin, and P. Gianiodis. 2005. Entrepreneurship and university-based technology transfer. *Journal of Business Venturing* 20: 241–263.
Marshall, A. 1923. *Industry and trade*. London: Macmillan.
McMillan, J., and C. Woodruff. 2002. The central role of entrepreneurs in transition economies. *Journal of Economic Perspectives* 16(3): 153–170.
Pennings, J. 1982. Organizational birth frequencies: An empirical investigation. *Administrative Science Quarterly* 27: 120–144.
Robinson, T., and E. Sexton. 1994. The effect of education and experience on self-employment success. *Journal of Business Venturing* 9: 141–156.
Say, J.-B. 1819. *Traite d'économie politique*. 4th edition. Trans. C. Prinsep, 1821. Boston: Wells and Lilly.
Schumpeter, J. 1911. *The Theory of Economic Development.* Trans. R. Opie, 1934. Cambridge, MA: Harvard University Press.
Shane, S. 2001. Technology opportunities and new firm creation. *Management Science* 47: 205–220.
Shane, S. 2003. *A general theory of entrepreneurship: The individual-opportunity nexus*. Northampton, MA: Edward Elgar.
Stuart, T., H. Huang, and R. Hybels. 1999. Interorganizational endorsements and the performance of entrepreneurial ventures. *Administrative Science Quarterly* 44: 315–349.
U.S. Bureau of the Census. 2004. *Statistical abstract of the United States: 2004–2005*. Washington, DC: Bureau of the Census.
Zahra, S., and W. Bogner. 2000. Technology strategy and software new ventures' performance: Exploring the moderating effect of the competitive environment. *Journal of Business Venturing* 15: 135–173.
Zietsma, C. 1999. Opportunity knocks – or does it hide? An examination of the role of opportunity recognition in entrepreneurship. In *Frontiers of entrepreneurship research*, ed. P. Reynolds et al. Babson Park: Babson College.

Entropy

Nicholas Georgescu-Roegen

A concept of momentous importance for our understanding of physical reality though it is, entropy is one of the most poorly understood even by many physicists as a keen



Entropy, Fig. 1

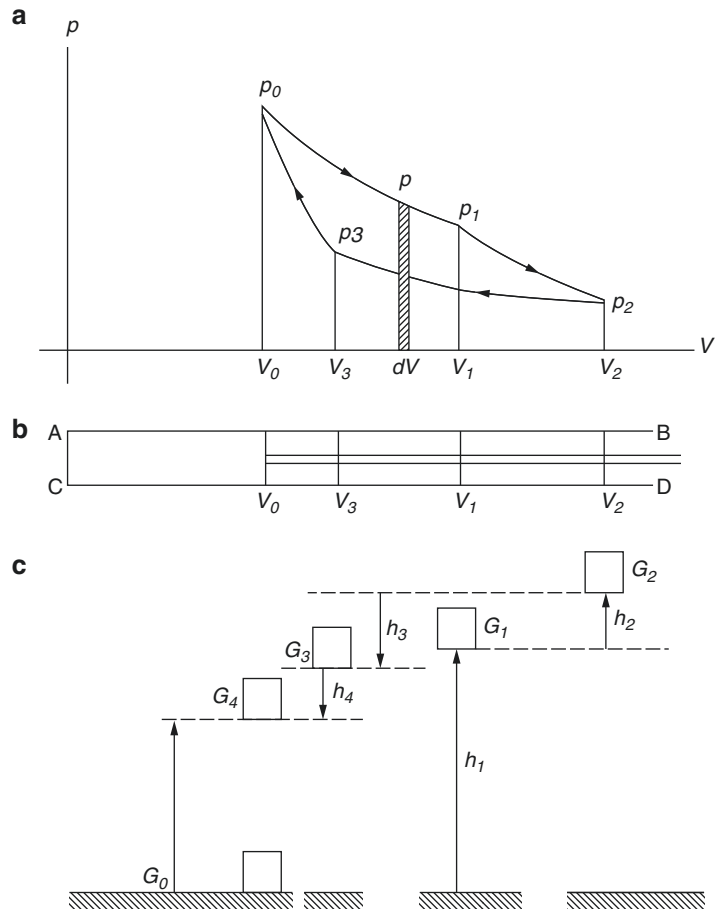
thermodynamicist, D. ter Haar, opined. A 'far-fetched' notion, 'obscure and difficult of comprehension' judged J. Willard Gibbs, the architect of statistical thermodynamics. It was apposite for Lord Snow to argue that some familiarity with the law of entropy, the second law of thermodynamics, separates the educated into two cultures. But this condition is quite curious given that the fountainhead of thermodynamics is anthropomorphic in a far more pronounced degree than that of any other branch of physics. No other physical concept belongs to our ordinary experience as inherently as heat and work or temperature and pressure. Indeed, thermodynamics is at bottom a physics of economic value as Sadi Carnot initiated it in his famous 1824 memoir about our efficient use of energy (Georgescu-Roegen 1971).

An explication of entropy at the ground level being beset with unusual difficulties, over the years the concept has received several, not always logically related, definitions. However, there is an original nature of that concept which can be grasped only by a punctilious description of the so-called Carnot cycle (Georgescu-Roegen 1987). This cycle – the pillar of thermodynamic theory – is a model of unmatched idealization of an engine that converts thermal into potential energy. It consists of a piston-and-cylinder (Fig. 1). The cylinder, ABCD, as well as the piston diaphragm, EF, are made of a perfect thermal insulator; the head of the cylinder, AC, instead is a perfect thermal conductor. The piston-rod turns a noncircular cam which raises or lowers a suspended weight, G. The space AEFC is filled with an ideal gas whose simple properties enable us to represent analytically the working of the engine (Fig. 2a). The model involves several other idealizations. First, the piston moves reversibly, i.e., with an infinitesimally slow speed, an assumption which does away with any friction. And since any such motion would require an infinite time, the point exposes one of the basic anthropomorphic limitations. Second, the piston performs no other work than turning the cam. And the cam on turning does not change its potential energy. Third, when the cycle begins, the volume of the gas is at, say, V_0, and its absolute temperature at T_1. A hot reservoir (a virtually limitless source of thermal energy at the same temperature T_1) is already attached at AC. On expanding, the gas absorbs thermal energy from the reservoir at a rate that keeps its temperature constant throughout. The work thus produced raises the weight from its initial position G_0. Because the internal energy of an ideal gas remains constant if its temperature does not change and because the work only raises the weight, the thermal energy, Q, absorbed by the gas from the reservoir must at any time be equal to the work, W.

Two observations call now for undivided attention. First, the phase considered so far proves that *it is possible to convert thermal energy (heat) from a single source into potential energy (work)*. Hence, in principle, we could sail by tapping the immense energy of the ocean water. Lord Kelvin (1882, I) was not exact therefore when he said in 1852 that '*It is impossible by means of an inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects*.' The true reason is not technical (as we have just seen), but still another anthropomorphic limitation: no human can work with a piston that expands beyond any limit; we must bring it back to begin another conversion. But to bring it back we need some energy. Of course, we could use the potential energy of the weight to push the piston back reversibly. That would not do for everything would be just as at the beginning.

There is, however, a technical means to bring the piston back by using less energy than Q. Sadi Carnot (p. 36) revealed the secret in a rather unnoticed law: *The fall of the caloric [thermal energy] produces more motive power [work] at inferior than at superior temperatures*. Let us then

Entropy, Fig. 2



assume that the gas expands only to V_1 which raises the weight to G_1 (Fig. 2). The performed work, W_1 ($=Q_1$), is represented by the area below the segment p_0p_1 of the curve $pV =$ constant which represents the isothermal expansion of the ideal gas. The reason for using a cam is that the work $p\mathrm{d}V$ is not uniform as V increases.

At V_1, an idealized operation takes place: instantaneously the hot reservoir is replaced by a perfect insulator covering AC and a different cam replaces the old. During this second phase the gas expands adiabatically (i.e. without exchanging heat with the outside) to V_2, converting Q_{12} of its internal energy into work, $W_{12} = Q_{12}$. This raises the weight further to G_2. By V_2 the temperature of the ideal gas has decreased to, say, T_2. Another idealized operation now replaces the thermal insulator with a cold reservoir of temperature T_2 and a different cam replaces the previous one. From V_2 on, Carnot's law applies: an energy Q_2 obtained by lowering G_2 to G_3 is transmitted to the reservoir through the intermediary of the gas which contracts so that its temperature and hence its energy remain constant. At V_3 again the cam is exchanged and the reservoir replaced by a perfect insulator over AC. The gas continues then to contract this time adiabatically, from V_3 to V_0. The work W_{34} supplied by the lowering of the weight from G_3 to G_4 is converted into internal energy $Q_{34} = Q_{12}$ because equal temperature changes cause equal changes of the internal energy of an ideal gas.

To sum up: the cycle of the piston is complete, from V_0 back to V_0. Yet the weight is not back at G_0, but at G_4. For the puzzling contrast we should recall the role of the cams in the model. It could

not be otherwise since the final work of the cycle, W, is positive, being represented by the area $p_0p_1p_2p_3$; accordingly,

$$W = Q_1 + Q_{12} - Q_2 - Q_{34} = Q_1 - Q_2 \quad (1)$$

The foregoing analysis proves the correctness of the so-called Carnot's principle, that any heat engine needs two sources of energy of different temperatures. The point is explicit in Planck's amended form of the second law of thermodynamics as given by Lord Kelvin:

> 'It is impossible to construct an engine which will work in a complete cycle, and produce no effect except the raising of a weight and the cooling of a heat-reservoir' (Planck 1897).

Planck thus negated what Wilhelm Ostwald called perpetual motion of the second kind.

From the equations of the curves of Fig. 2a there follows an expression of Carnot's law

$$Q_2/T_2 - Q_1/T_1 = 0. \quad (2)$$

This surprising relation led Rudolf Clausius (in an 1854 essay) to interpret Q/T as *the transformation equivalent-value* of Q at the temperature T. And, as Lord Kelvin had done a few months earlier, he proved that for a reversible cycle

$$N = \sum Q_i/T_i = 0 \text{ or } N = \int \mathrm{d}Q/T = 0. \quad (3)$$

Both based the *theorem* on the first law of thermodynamics, $W = \sum Q/T = 0$ for an isolated system, and on a formulation of the second law. Clausius', which is equivalent to that of Kelvin-Planck, is still the most transparent one: *Heat cannot by itself pass from a colder to a warmer body.*

Clausius also proved the epochal result that, with the convention for signs as in (2), for an *irreversible* cycle

$$\int \mathrm{d}Q/T > 0 \quad (4)$$

In 1865, on the basis of (3) he defined a new thermodynamic function, S, by

$$S_a - S_b = \int_a^b \mathrm{d}Q/T \quad (5)$$

for any reversible path from a state a to b. Then, by (3) and (4), always

$$\Delta S \geq \int \mathrm{d}Q/T, \quad (6)$$

the equality prevailing only for reversible transformations.

At that juncture, Clausius replaced the term, 'transformation' by its Greek equivalent, *entropi*, and ended the memoir by the famous stanza:

1. *The energy of the universe is constant.*
2. *The entropy of the universe tends to a maximum.*

Although S was then defined only for thermal equilibrium, it was one of the greatest novelties ever thought up by a scholarly mind.

Once the idea that heat is not an indestructible fluid but a kind of molecular motion gained credence, to explain thermodynamic phenomena by the laws of mechanics became a vital programme. The most ambitious attempt was that of an 1872 epochal, albeit hard-going, essay by Ludwig Boltzmann (Brush 1966). He argued that if the collisions between molecules follow some (apparently innocuous) rules, the distribution, $f(x,t)$, of their velocities at any time, t, is such that

$$H(t) = \int f \log f \, \mathrm{d}x \quad (7)$$

is never increasing,

$$\mathrm{d}H(t)/\mathrm{d}t \leq 0, \quad (8)$$

the equality prevailing only for thermal equilibrium. Naturally, Boltzmann went on to claim not only that $S = -H$, but that (7) defined entropy for any thermodynamic system as well.

Since to derive an irreversible property from a completely reversible axiomatic basis was as incredible a feat as claiming that the angles of a triangle in an Euclidean plane add to more than two right angles, protests had to come. In 1876, a Boltzmann colleague, Joseph Loschmidt, pointed out that if at any time the velocities of a system satisfying (8) are reversed, we obtain a system for which H increases. And twenty years later, Ernst Zermelo, a pupil of Max Planck, recalled that in 1890 Henri Poincarè had proved that any finite mechanical system must eventually return as close as we wish to any of its *previous* positions (cf. the elementary illustrations in Georgescu-Roegen 1971, pp. 154f). Hence, if H decreases for some time, it must necessarily increase before the return. Boltzmann was at a loss to defend his strong position of 1872 (see Brush 1966). He asserted that if velocities are reversed the entropy most probably would still increase, since his theorem was not based on 'the nature of the forces' but on the immensely greater probability of the initial conditions that yield (8). But there is no reason for these conditions – known as *Stossenanzahl* (statistical postulate) – to be perpetuated from one collision to the next (Georgescu-Roegen 1971, App. C). As to Zermelo's entropy recurrence, Boltzmann dismissed it as irrelevant in practice because of the extremely long time for the return of present conditions (Brush 1966). Loschmidt's objection, however, had a weak point, namely, that systems with reversed velocities do not exist always in actuality. And in favour of Zermelo's, one could imagine that we may be in the middle of a recurrence period begun eons ago.

Pressed by persistent criticism Boltzmann abandoned his purely mechanistic basis of entropy, to concede that his H-theorem 'can never be proved from the equations of motions alone, [it] can only be deduced from the laws of probability' (Boltzmann 1895). Already in 1877, he anchored the law of increasing entropy on the postulate that every state always passes to one of greater probability. Reasonable though the postulate may seem, it is not supported by probability laws: the occurrence of highly improbable bridge hands is subject to no sequential condition (Georgescu-Roegen 1971, VI. 1 and App. F). Then, observing that any state ultimately reaches one of the molecular structures corresponding to thermal equilibrium, and that the number, Ω, of those structures is far greater than that of any other state, Boltzmann proposed that for thermal equilibrium

$S = \log \Omega$, which after the dimensional correction by Max Planck became

$$S = k \log \Omega, \tag{9}$$

where k is now registered as Boltzmann's constant. As has been observed by Planck (1897), the logarithm function must be used in (9) because for a thermal equilibrium composed of two such states we know that $S = S_1 + S_2$ whereas $\Omega = \Omega_1 + \Omega_2$. Later, Boltzmann (1896/8) connected (9) with (7) and (8) by observing that n_i, being the number of molecules in some state i, with $n = \sum n_i$, then the number of possible structural combinations for that system is

$$W = n!/n_1!n_2! \ldots n_m! \tag{10}$$

Hence, granting that every n_i is sufficiently great, by Stirling's asymptotic formula for $n!$ we obtain $\log W = \Sigma n_i \log(n_i/n)$, where by putting $f_i = n_i/n$, by analogy to (7) and (9) Boltzmann put

$$S = -nk\Sigma f_1 \log f_1 = -nkH, \tag{11}$$

still another expression for entropy. At least, this formula fared well with the advent of Planck's quantum theory, which explains why Planck, who at first opposed Boltzmann's position, ultimately changed his mind (1897, seventh edn).

To buttress the probabilistic interpretation of entropy, Boltzmann (1896/8) ultimately brought in the ideas of ordered and disordered states. Herman von Helmholtz had already defined (in 1892) entropy as the measure of disorder, an unfortunate connection as disorder certainly cannot be defined analytically. Curiously, this definition has had amazing success: the entry for 'entropy' in *The Encyclopedia of Philosophy* opens with it.

Because the endeavours to systematize the various ways of looking at entropy could not arrive at

a simple, natural explanation of that notion, writers have moved deeper and deeper into purely formal lucubrations, so that Jacques Hadamard in his 1910 review of J. Willard Gibbs's treatise could say that statistical thermodynamics is only mathematics (cited in Georgescu-Roegen 1971). Yet as early as 1882 Helmholtz placed entropy in an accessible cadre as he showed that the internal energy, U, of an isolated system consists of free (or available) energy, F, and of bound (or unavailable) energy measured by TS. Hence, if we represent U by a rectangle of base T, the entropy would indicate the height for separating the *bound* energy. It was with the introduction of the relation $F = U - TS$ that the importance of the singular concept, entropy, was brought to the surface. It was only thereafter that one could translate Clausius's entropy law into

> The free energy of any isolated system continuously degrades into bound energy.

It is this form that pinpoints the reason for the supreme role of entropy in all nature, as has been recognized by great luminaries such as Sir Arthur Eddington and Albert Einstein. Interestingly, Lord Kelvin who first spoke (in 1852) of 'The Universal Tendency in Nature to the Dissipation of Energy', hardly ever used the term 'entropy'. And Walter Nernst, another illuminator of thermodynamics, even decided not to have recourse to it.

We should next recall that just as the foundation of classical thermodynamics was being laid, Herbert Spencer came forth with some tenets that presaged Darwin's own theory. One was that 'the homogeneous is the hotbed of the heterogeneous', which looks like a characterization of living systems. Lord Kelvin (quoted earlier) as well as Helmholtz thus were not prepared to admit that the entropic degradation applies to animated matter, too. Later Henri Bergson even claimed that life opposes that degradation. Of course, life does not violate the entropy law, for as Erwin Schrödinger was to put it not very long ago, a living creature is not an isolated system: it exchanges entropy with its environment. Yet from the fact that such a phenomenon is not impossible thermodynamically, it does not follow at all that it should also exist. That is why several scholars have argued that in nature there must also be at work an anti-entropy principle – *ektropy*, coined by G. Hirth and adopted by Felix Auerbach, or *anti-chance*, Sir Arthur Eddington's term (Georgescu-Roegen 1971). The dissipative structures, recently set forth by Ilya Prigogine (1980) to portray the entropic process specific to living organisms imply as Prigogine recognizes, a new crowning of the Spencerian tenet.

A real imbroglio involving the entropy concept grew from the seminal work of Claude H. (Shannon and Weaver 1948) on the purely technical problem of communication, which is to find out how many distinct sequences (messages) of a given length can be formed by a code, a set of different single signals. For the communication engineer it is totally irrelevant what each message may mean. However, that meaning must be understood by both the originator and the intended receiver. Knowledge thus passes from the former to the latter; while in transit, it is information (a distinction analogous to the heat being thermal energy in transit).

For messages transmitted in a vernacular language, Shannon found that the ratio of messages per letter is given by the same formula as $-H$ in (11) if the f_i's represent the statistical frequencies of the corresponding alphabet. Seeking a name for his new formula (a measure of the efficiency of the code), Shannon accepted the suggestion of John von Neumann: 'entropy'. But the mere coincidence of formulae was not a basis for justifying the terminological transfer. We would not call 'kinetic energy' the second moment of a distribution just because both formulae are a sum of squares: $\sum \eta_i X_i^2$. With that transfer, the concept of entropy started travelling from one domain to another with hardly any discrimination. One now speaks of the income distribution in country A being greater than in B, although the student interested in B will certainly think otherwise. And the reader is dizzied by the frequent phrases in which 'information', 'knowledge', 'negentropy', intervene pell-mell. Further, O, Onicescu suggested that $\sum f_i^2$ can also serve as 'informational energy', which, of course, could not be related to the real entropy. Most important, in a

vignetted article, 'The Bandwagon', Shannon himself protested without delay the use of 'entropy' beyond the domain of technical communication (Georgescu-Roegen 1975).

The idea that the entropic degradation of anything is in fact a 'loss of information' was set forth earlier by a consummate thermodynamicist, G.N. Lewis (*Science*, 1930). But it was E.T. Jaynes who after the spread of Shannon's theorem set out to erect thermodynamics on that basis alone. In spite of its bizarreness, or perhaps because of it, the idea is still running in some circles. So, in his recent primer (*The Second Law*, 1984), P.W. Atkins was in good order to deliberately omit any reference to entropy and information because of the 'muddleheadedness' of the idea that entropy is not a property, of an engine but of the engineer's mind.

While the concept of entropy was thus converted almost at will, a vital issue found no place in thermodynamics, namely, the macroscopic role of matter. Matter is mentioned but only indirectly, as friction. Prigogine (1955) did extend the domain of thermodynamics from closed (impermeable to matter) systems to open systems, but he considered matter only as a vehicle of energy – the heat carried by a red-hot iron, for instance. No one seems to have derived the important object lesson from Gibbs's proof that the interdiffusion of two gases of the *same* temperature increases entropy. The increase is due to the entropic degradation of matter. To fill this lacuna, a new law of thermodynamics states that

Perpetual Motion of the Third Kind is Impossible

Which means that no closed (not to be confused with 'isolated') system can perform work indefinitely at constant rate. The reason is that macroscopic matter also degrades entropically (Georgescu-Roegen 1980).

With its exotic name and its complicated fate even within the evolution of thermodynamics, entropy has become a word of great alluring power. Occasionally, littérateurs have used it manifestly as a selling point, 'Entropy' by Thomas Pynchon in 1960, *Against Entropy* by Michael Frayn (1967). Clausius certainly did not foresee this development from his coinage of the bizarre word.

See Also

▶ Information Theory

Bibliography

Boltzmann, L. 1895. On certain questions of the theory of gases. *Nature* 51: 413–415.

Boltzmann, L. 1896–98. *Lectures on gas theory*. Trans. from German by S.G. Brush. Berkeley: University of California Press, 1964.

Brush, S. 1966. *Kinetic theory*, vol. 2. Oxford: Pergamon Press.

Carnot, S. 1824. *Reflections on the motive power of fire, and on the machines fitted to develop that power*. Trans. and ed. R.H. Thurston. New York: Dover, 1960.

Clausius, R. 1867. In *The mechanical theory of heat with its application to the steam engine and to the physical properties of bodies*, ed. T.A. Archer. London: John van Voorst.

Georgescu-Roegen, N. 1971. *The entropy law and the economic process*. Cambridge, MA: Harvard University Press.

Georgescu-Roegen, N. 1975. The measure of information: A critique. In *Proceedings of the third International Congress of Cybernetics and Systems, Bucharest, August 25–29 1975*, ed. J. Rose and C. Bilciu. New York: Springer-Verlag.

Georgescu-Roegen, N. 1980. Matter: A resource ignored by thermodynamics. In *Future sources of organic raw materials*, ed. L.E. St-Pierre and G.R. Brown. Oxford: Pergamon Press.

Georgescu-Roegen, N. 1987. *The promethean destiny of mankind's technology*. Brighton: Wheatsheaf.

Planck, M. 1897. *Treatise on thermodynamics*. Translated from the 7th German edn by A. Ogg. New York: Dover, 1925.

Prigogine, I. 1955. *Thermodynamics of irreversible processes*, 3rd ed. New York: Interscience Publishers, 1967.

Prigogine, I. 1980. *From being to becoming*. San Francisco: W.H. Freeman.

Rankine, W.J.M. 1881. In *Miscellaneous scientific papers*, ed. W.J. Millar. London: Charles Griffin.

Shannon, C., and W. Weaver. 1948. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Thomson, W., and Baron Kelvin. 1882. *Mathematical and physical papers*. Cambridge: Cambridge University Press.

von Helmholtz, H. 1882. On the thermodynamics of chemical processes. In *Physical memoirs*, vol. I, ed. H. von Helmholtz. London: Taylor & Francis, 1888–90.

Entry and Market Structure

B. Curtis Eaton

Entry – and its opposite, exit – have long been seen to be the driving forces in the neoclassical theory of competitive markets. Long-run equilibrium in such a market requires that no potential entrant finds entry profitable, and that no established firm finds exit profitable. In conjunction with the price-taking assumption, the first condition requires that price be no greater than minimum average cost, $\underline{AC}$, and the second that price be no less than $\underline{AC}$. Hence, in equilibrium, price is equal to $\underline{AC}$. There is very little more to the theory of equilibrium in a competitive market than this simple yet powerful story of no-entry and no-exit.

Surprisingly, considerations of entry and potential competition, so central to the economist's view of competitive markets, played almost no role in oligopolistic and monopolistic markets until the work of Bain (1956) and Sylos-Labini (1956) in the mid–1950s. One important strand of the modern literature on entry combines, in essence, their insights into the role of potential competition in oligopolies with Schelling's (1956) ideas on commitment. This essay focuses on this strand of literature.

In reviewing Bain (1956) and Sylos-Labini (1956), Modigliani (1958) proposed what has come to be known as the *limit-output* (more commonly, *limit-price*) model, which is a formalization of one of the key ideas in these books. One version of this model has been the focal point of much of the recent literature on entry. Consider a market for some undifferentiated good, currently served by *one established firm*, in which demand and cost conditions are unchanging over an infinite time horizon. Now suppose that all potential entrants take the established firm's output today, denoted by X, as the output which it will produce tomorrow and forever – the so-called Sylos postulate. Let $\underline{X}$ denote the smallest value of X such that the maximized profit of a representative potential entrant is non-positive. This value, $\underline{X}$, is called the *limit output* (and the corresponding price the *limit price*) because, given the Sylos postulate, there will be no entry if and only if $X \geq \underline{X}$. (We assume for convenience that zero profit does not induce entry.)

Two important insights are already clear. First, potential competition constrains the ability of the established firm to exploit its position of market power since the no-entry condition ($X \geq \underline{X}$) places a lower bound on industry output in long-run equilibrium. Second, by producing at least $\underline{X}$ units of output, the established firm can deter entry. In one case central to the evolution of the literature, entry deterrence is also always profitable.

Suppose that the established firm's average cost function is nowhere upward-sloping, and that the interest rate is not too high. In this situation, the established firm will always choose to deter entry by producing at least $\underline{X}$. There are two sorts of solutions. If the ordinary monopoly output, M, is greater than or equal to $\underline{X}$, then the monopolist will produce M, and we have what we could call *natural monopoly*, in the positive sense of the term. If $M < \underline{X}$, the monopolist produces X to deter entry strategically, a case of *artificial monopoly*. When $M < \underline{X}$ the monopolist must choose between deterring and accommodating entry. Relative to the deterrence strategy – produce $\underline{X}$ today and forever – accommodation produces larger profit for the one established firm today (since today's output will be M) but smaller profit tomorrow and forever (since price will be no higher than the limit price and the established firm's output will be less than $\underline{X}$). If the rate of interest is not too high, it is obvious that the established firm will choose to deter entry.

The essence of both solutions is a message which the established firm wants to communicate to potential entrants. 'If you enter, then my output will be (no smaller than) $\underline{X}$.' This 'deterrence message' has the property that it deters entry – if it is believed. It raises the obvious credibility question, 'Would the established firm really produce the promised output post-entry'? Much of the recent literature on entry has implicitly or

explicitly focused on this question. Four interrelated insights have emerged.

First, to answer the credibility question we can use Schelling's distinction between threats and commitments (1956). The deterrence message is a commitment if, entry having occurred, it is in the established firm's self-interest to produce $\underline{X}$, or if the production of $\underline{X}$ follows automatically. Otherwise the message is a threat and is not credible. To implement this approach we need a model of oligopoly. Given such a model, all interested parties can compute the post-entry equilibrium, providing a direct answer to the credibility question.

Second, by virtue of being there first, the established firm has the opportunity to make *irreversible decisions* which alter its real economic circumstances in any post-entry oligopoly game. These irreversible decisions can sometimes make the established firm a more aggressive competitor post-entry, and therefore serve to make the deterrence message more believable. That is, the established firm has the opportunity to do some things prior to entry which cannot be undone subsequent to entry, and which affect the profitability of entry. It is convenient to refer to these irreversible decisions as *commitments*. As Spence (1977) observed, since the rate at which output is produced is reversible, producing the limit output prior to entry is not a commitment and therefore has no bearing on the credibility of the deterrence message. However, holding the capacity or capital to do so is – provided that it is specific. By acquiring specific capital, the established firm reduces its marginal cost, making it more aggressive in any post-entry oligopoly game. Inventory (analysed by Ware 1985) is a particularly illuminating commitment since it puts the firm in the position of having zero marginal cost post-entry (until its inventory is exhausted).

The third insight, which arises from the attempt to implement the first two, concerns the *form* of the game which established firms and potential entrants play, and the appropriate *equilibrium concept* which this form seems to imply. Even in the simplest of circumstances, any game involving established firms and potential entrants is a *multi-stage game* played out in real time with two important features: (1) commitments, which inevitably involve *sunk costs*, are made in earlier stages of the game; (2) the net revenues which justify these sunk costs are generated only in later stages. As Brander and Spencer (1983) observe, this form is an unavoidable feature of economic reality. Product development costs, for example, must be sunk prior to production; the costs associated with specific capital goods must be sunk prior to production, and so on.

A rational firm must therefore think of commitments in the way it thinks of other investment decisions. In particular, it must form expectations about how decisions with respect to today's commitments will alter its net revenues tomorrow, and thereafter. In the presence of sunk costs, *rational* or *consistent* expectations are a desirable feature of any equilibrium concept. If firms' expectations are not constrained to be rational, then any outcome is possible – that is, given an outcome, there is a set of expectations which will produce it. Rational expectations are then necessary to constrain results. See, for example, the discussions in Eaton and Lipsey (1979, 1980) and Dixit (1980) on this point. Given this view, Selten's (1975) notion of sub-game perfection is the appropriate equilibrium concept in these entry games.

To convey the flavour of modern theories of entry and to see how rational expectations enter the analysis, it is instructive to write down a simple entry game and to consider the way in which one finds the perfect equilibrium. Most of the recent literature on entry has focused on exercises which involve one established firm and one potential entrant, and which are not a great deal more complex than the following illustration.

Consider an entry game played in three stages. In stage 1, firm 1 (the established firm) chooses the value of some commitment, c_1. In stage 2, knowing the value of c_1, which firm 1 chose in stage 1, firm 2 chooses a value for its commitment, c_2. By appropriate choice of units, we can interpret c_1 and c_2 as costs, which once they are incurred are sunk. These sunk costs might, for example, be expenditures on advertising or on cost-reducing research and development. In stage 3, the two firms play a market game in

which goods are produced and sold and the net revenues which justify the upstream sunk costs are realized.

To find the perfect equilibrium of this game we work backwards.

Stage 3 Given an oligopoly model which determines the equilibrium of the market game, the net revenue to each firm in stage 3 is determined by c_1 and c_2. Denote these net-revenue functions by $\prod_1(c_1, c_2)$ and $\prod_2(c_1, c_2)$ If, for example, the oligopoly model is the Cournot model, then in stage 3 each firm chooses its own quantity to maximize its revenues minus its *avoidable* costs. The net-revenue functions are simply revenues minus avoidable costs in the Cournot equilibrium.

Stage 2 If firm 2 is to have rational expectations, it must know $\prod_2(c_1, c_2)$. This, of course, means that it knows the oligopoly model which determines the equilibrium of the market game. In stage 2, knowing its net-revenue function and the value of c_1, firm 2 chooses c_2 to maximize $[\prod_2(c_1, c_2) - c_2]$. The solution to this maximization problem determines c_2 as a function of c_1 : $c_2 = g(c_1)$.

Stage 1 Rational expectations for firm 1 means that it knows both $g(c_1)$ and $\prod_1(c_1, c_2)$. In stage 1 it chooses c_1 to maximize $[\prod_1(c_1, c_2) - c_2]$ subject to $c_2 = g(c_1)$. The only endogenous variable in this maximization problem is c_1 and the solution to it therefore determines a value for c_1 say c^*_1. Firm 2 then chooses $\Pi_1(c^*_1, c^*_2)$, and in the third stage of the game the firms realize $\Pi_2(c^*_1, c^*_2)$ and $c^* = g(c^*)$.

In this sort of game the established firm may or may not be able to deter entry, and if it is able, it may or may not choose to. Duopoly solutions will, however, be asymmetric. The established firm will rig the duopoly market structure to its own advantage.

Using this approach, or one that is in the spirit of this one, the recent literature on entry has focused on many of the commitments which established firms can and, indeed, must make. Advertising (Cubbin 1981), brand proliferation (Schmalensee 1978), the location of retail outlets (Eaton and Lipsey 1979), patenting (Gilbert and Newbery 1982), learning-by-doing (Spence 1981), the durability of specific capital (Eaton and Lipsey 1980), the exercise of monopsony power (Salop and Scheffman 1983) and, of course, specific capital (Spence 1977; Dixit 1980 and Ware 1984) are just some of the vehicles for commitment which have been considered.

This rich set of games and possible solutions brings us to the fourth insight in this literature. Implicit in this way of thinking about oligopolistic markets, and the role which entry plays in those markets, is much more than a theory of how one established firm strategically positions itself with respect to one potential entrant. There is, in this paradigm, a theory of market structure, a theory which remains largely unexplored.

See Also

- ▶ Contestable Markets
- ▶ Limit Pricing
- ▶ Natural Monopoly
- ▶ Predatory Pricing

Bibliography

Bain, J.S. 1956. *Barriers to new competition*. Cambridge: Harvard University Press.

Brander, J.A., and B.J. Spencer. 1983. Strategic commitment with R&D: The symmetric case. *Bell Journal of Economics* 14(1): 225–235.

Cubbin, J. 1981. Advertising and the theory of entry barriers. *Economica* 48: 289–298.

Dixit, A. 1980. The role of investment in entry deterrence. *Economic Journal* 90: 95–106.

Eaton, B.C., and R.G. Lipsey. 1979. The theory of market preemption: The persistence of excess capacity and monopoly in growing spatial markets. *Economica* 46(182): 149–158.

Eaton, B.C., and R.G. Lipsey. 1980. Exit barriers are entry barriers: The durability of capital as a barrier to entry. *Bell Journal of Economics* 11(2): 721–729.

Gilbert, R.J., and D.M.G. Newbery. 1982. Preemptive patenting and the persistence of monopoly. *American Economic Review* 72(3): 514–526.

Modigliani, F. 1958. New developments on the oligopoly front. *Journal of Political Economy* 66: 215–232.

Salop, S.C., and D.T. Scheffman. 1983. Raising rivals' costs. *American Economic Review* 73(2): 267–271.

Schelling, T.C. 1956. An essay on bargaining. *American Economic Review* 46: 281–306.
Schmalensee, R. 1978. Entry deterrence in the ready-to-eat breakfast cereals industry. *Bell Journal of Economics* 9(2): 305–327.
Selten, R. 1975. Re-examination of the prefectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4(1): 25–55.
Spence, A.M. 1977. Entry, capacity, investment and oligopolistic pricing. *Bell Journal of Economics* 8(2): 534–544.
Spence, A.M. 1981. The learning curve and competition. *Bell Journal of Economics* 12(1), Spring: 49–70.
Sylos-Labini, P. 1956. *Oligopolio e progresso technico*. Milan: Giuffre.
Ware, R. 1984. Sunk costs and strategic commitment: A proposed three-stage equilibrium. *Economic Journal* 94: 370–378.
Ware, R. 1985. Inventory holding as a strategic weapon to deter entry. *Economica* 52: 93–101.

Envelope Theorem

Eugene Silberberg

Abstract

The envelope theorem appeared in economics following the 1931 Viner–Wong diagram (incorrectly drawn in the original paper). This famous paper indicated that, starting at some minimum cost input combination, the change of average cost when output changed was the same whether or not other inputs were allowed to vary or were held fixed. This puzzling result remained mostly a curiosity until the 1970s when, with the use of a generalization of this diagram, the modern theory of duality was developed. This new approach to comparative statics provided a clearer explanation for the appearance of refutable implications in maximization models.

Keywords

Comparative statics; Constrained optimization models; Cost functions; Envelope theorem; Lagrange multipliers; Primal-dual model; Shadow pricing

JEL Classifications

D2

The origin of this famous theorem is the discussion between Jacob Viner (1931) and his draughtsman Y.K. Wong concerning the relationship between short- and long-run average cost curves. Viner had apparently reasoned that since in the long run average costs should be at a minimum, the long-run average cost (LRAC) curve should not only always be below the short-run average cost (SRAC) curves, but should also pass through the minimum points of each short-run curve. Wong pointed out the impossibility of this joint occurrence, and Viner opted to draw the long-run curve through the minimum points, thereby necessarily passing above sections of the short run curves. It was also puzzling (in the now corrected diagram) that at the point of tangency between the LRAC and a SRAC, the rate of change of average cost with respect to output was the same when capital was fixed as when it was allowed to vary. The puzzle was solved by Samuelson (1947), who showed in a general way why the long-run curve would be the 'envelope' curve to the set of short-run curves. Perhaps the most surprising result of all was that this seeming mathematical curiosity turned out to be the fundamental basis for the development of refutable comparative statics implications in economics.

Unconstrained Maximization Models

The most general comparative statics model with explicit maximizing behaviour is *maximize* $y = f(x, \alpha)$ subject to $g(x, \alpha) = 0$, where $x = (x_1,..., x_n)$ is a vector of decision variables, $\alpha = (\alpha_1,..., \alpha_m)$ is a vector of parameters (though for simplicity, we treat a as a scalar in the discussion below), and g(·) represents one or more constraints. Models at this level of generality, however, imply no refutable implications and are hence largely uninteresting. In particular, there are never refutable implications for parameters that enter the constraint (see, for example, Silberberg and Suen 2000). We therefore initially restrict the

analysis to models of unconstrained maximization:

$$maximize\ y = f(x, \alpha) \tag{1}$$

The necessary first-order conditions (NFOC) are

$$f_i(x, \alpha) = 0\ \ i = 1, ..., n \tag{2}$$

The sufficient second-order conditions (SSOC) are that the *Hessian* matrix $\mathbf{H} = (f_{ij})$ is negative definite. Alternatively, the principal minors of order (size) k of the Hessian determinant $H = |f_{ij}|$ have sign $(-1)^k$. Assuming the sufficient second-order conditions hold, we can in principle 'solve' for the n explicit choice functions $x = x^*(\alpha)$. Of course, since these choice functions are the result of solving the NFOC simultaneously, each individual x_i is a function of *all* the parameters, not just ones which might appears in some f_i.

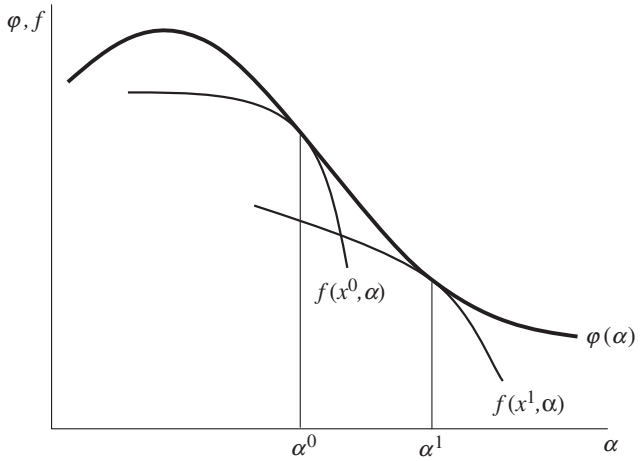
Substituting the x_i^*'s into the objective function yields the *indirect objective function* $\phi(\alpha) = f(x^*(\alpha), \alpha)$, the maximum value of f for given α. Since $\phi(\alpha)$ is by definition a maximum value, $\phi(\alpha) \geq f(x, \alpha)$, but $\phi(\alpha) = f(x, \alpha)$ when $x = x^*$. In Fig. 1, a typical $\phi(\alpha)$ is plotted. For an arbitrary α^0, an $x^0 = x^*(\alpha^0)$ is implied. Consider the behaviour of $f(x, \alpha)$ when the x_i's are held fixed at x^0 as opposed to when they are variable. When $\alpha = \alpha^0$, the 'correct' x_i's are chosen, and therefore $\phi(\alpha) = f(x^0, \alpha)$ at that one point. However, both to the left and to the right of α^0, the 'wrong' (that is non-maximizing) x_i's are chosen, and, since $\phi(\alpha)$ is the *maximum* value of f for given α, $f(x^0, \alpha) \leq \phi(\alpha)$ in any neighbourhood around α^0. This implies that ϕ and f must be tangent at α^0 - (assuming differentiability), and, moreover, f must be either more concave or less convex than ϕ there. Since this must happen for arbitrary α, similar tangencies occur at other values of α. It is apparent from Fig. 1 that $\phi(\alpha)$ is the *envelope* of the $f(x_1, x_2, \alpha)$'s for each α. What surprised most researchers was the discovery that all comparative statics theorems in maximization models are in fact consequences of the relative curvatures of ϕ and f.

From the above discussion, the function $F(x, \alpha) = f(x, \alpha) - \varphi(\alpha)$ has a maximum of zero, with respect to both x and α. Thus we consider the *primal–dual* model

$$maximize\ F(x, \alpha) = f(x, \alpha) - \phi(\alpha) \tag{3}$$

where the maximization runs over x and also α. (In the latter instance, we ask, for given x_i's, what values of the parameters would make these x_i's the maximizing values?) The NFOC with respect to x are the same as in the original model. With respect to α, the NFOC yield the famous 'envelope theorem' which is the tangency of f and ϕ in Fig. 1:

$$F_\alpha = f_\alpha - \phi_\alpha = 0 \tag{4}$$



Envelope Theorem, Fig. 1

In the α dimensions, the second-order conditions are simply

$$F_{\alpha\alpha} = f_{\alpha\alpha} - \phi_{\alpha\alpha} \leq 0. \tag{5}$$

This inequality says that in the α dimensions, f is relatively more concave than ϕ. (When α is a vector, this second-order condition is that the Hessian matrix ($F_{\alpha\alpha}$) is negative semi-definite.)

This is the fundamental geometrical property that underlies all comparative statics relationships. The NFOC (4) are identities when $x = x^*$. That is,

$$\phi_\alpha(\alpha) \equiv f_\alpha(x^*(\alpha), \alpha) \tag{6}$$

Differentiating with respect to α,

$$\phi_{\alpha\alpha} \equiv \sum_1^n f_{\alpha i} \frac{\partial x_i^*}{\partial \alpha} + f_{\alpha\alpha} \tag{7}$$

Rearranging terms, using (5) and invariance to the order of differentiation,

$$\phi_{\alpha\alpha} - f_{\alpha\alpha} \equiv \sum_1^n f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0 \tag{8}$$

This is the fundamental relation of comparative statics. From it, we can derive Samuelson's famous 'conjugate pairs' theorem that refutable implications occur in maximization models when and only when a parameter enters one and only one firstorder condition. For in that case, where say α enters only $f_i = 0, f_{j\alpha} = 0, j \neq i$, and so (8) reduces to one term:

$$f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0 \tag{9}$$

In this case we can say that the response of x_i is in the same direction as the disturbance to the equilibrium (or, in the case of minimization models, in the opposite direction). For example, consider the profit-maximization model

$$\begin{aligned} \textit{maximize } \pi &= f(x, w, p) \\ &= p\theta(x_1, \ldots, x_n) - \sum w_i x_i \end{aligned}$$

Each parameter w_i enters only the ith NFOC, and $f_{x_i w_i} = -1$, so that (9) yields the slope property $\partial x_i^* / \partial w_i \leq 0$; the factor demand functions are downward sloping in their own price.

The envelope theorem also yields the non-intuitive 'reciprocity' conditions. Suppose there are two parameters α and β. Then from invariance of second partial derivatives to the order of differentiation (Young's theorem), $\phi_{\alpha\beta} = \phi_{\beta\alpha}$. Using Eq. (6) above,

$$\sum f_{i\alpha} \frac{\partial x_i^*}{\partial \beta} = \sum f_{i\beta} \frac{\partial x_i^*}{\partial \alpha} \tag{10}$$

When the objective function contains a linear expression such as in the profit maximization model, that is, $w_1 x_1 + \ldots + w_n x_n$, we have $f_{x_i w_i} = -1$ and $f_{x_i w_i} = 0, i \neq j$. In that case, (11) reduces to the simple expression $\frac{\partial x_i^*}{\partial w_j} = \frac{\partial x_j^*}{\partial w_i}$. This result also occurs in consumer theory for the Hicksian demands.

Constrained Maximization Models

Consider now the general comparative statics model with constraints, *maximize* $y = f(x, \alpha)$ subject to $g(x, \alpha) = 0$, where $g(\cdot)$ represents one or more constraints. Assuming just one constraint for the moment, the Lagrangian for this model is $L = f(x, \alpha) + \lambda g(x)$, producing the NFOC

$$L_i = f_i(x, \alpha) + \lambda g_i(x, \alpha) = 0 \; i = 1, \ldots, n \tag{11}$$

$$L_\lambda = g(x, \alpha) = 0 \tag{12}$$

Assuming the SSOC, we can in principle 'solve' for the $n + 1$ explicit choice functions $x = x^*(\alpha)$ and $\lambda^*(\alpha)$. We derive the indirect objective function as before by substituting the x_i^*'s into the objective function producing $\phi(\alpha) = f(x^*(\alpha), \alpha)$, the maximum value of f for given α, now also subject to the constraint. Proceeding as above, since $\phi(\alpha)$ is by definition a maximum value, $\phi(\alpha) \geq f(x, \alpha)$, but $\phi(\alpha) = f(x, \alpha)$ when $x = x^*$. Thus the function $F(x, \alpha) = f(x, \alpha) - \phi(\alpha)$ has a (constrained) maximum of zero, with respect to both x and α. Thus we consider the *primal–dual* model

$$\text{maximize } F(x,\alpha) = f(x,\alpha) - \phi(\alpha) \tag{13}$$

$$\text{subject to } g(x,\alpha) = 0 \tag{14}$$

where the maximization runs over x and also α. The Lagrangian for this model is

$$L = f(x,\alpha) - \phi(\alpha) + \lambda g(x,\alpha) \tag{15}$$

The first-order conditions with respect to x are the same as in the original model. With respect to α, we get the envelope theorem in its most general form,

$$\phi_\alpha = L_\alpha = f_\alpha + \lambda g_\alpha \tag{16}$$

At this level of generality, it is not possible to generate any useful curvature properties of $\phi(\alpha)$. However, consider the case where α does not enter any constraint. In that case, $g_\alpha \equiv 0$ and the NFOC reduce to (4) above, that is, $F_\alpha = f_\alpha - \phi_\alpha = 0$. Moreover, when a does not enter the constraint, the primal–dual model is *an unconstrained maximization in* α. Hence in the α dimensions, the second-order conditions are as before:

$$F_{\alpha\alpha} = f_{\alpha\alpha} - \phi_{\alpha\alpha} \leq 0. \tag{17}$$

Thus in this important class of models, the comparative statics are identical to the models with no constraints. We obtain the inequalities (8) and (9) in the same manner as above.

Consider now an important class of models having the structure maximize $f(x)$ subject to $g(x) = k$, where we suppress all parameters except k, which is the focus of this analysis. The Lagrangian for this model is $L = f(x) + \lambda(k - g(x))$; assuming the NFOC and SSOC are valid, we solve for the explicit choice functions $x = x^*(k)$ and $\lambda^*(k)$. The indirect objective function is $\phi(k) = f(x^*(k))$, the maximum value of f for given k. The envelope theorem (16) yields

$$\phi_k = \lambda^*(k) \tag{18}$$

Suppose the function f represents the value of output, and the constraint describes a limitation on that value due to the scarcity of some resource, measured by the value of k. Then the Lagrange multiplier imputes a 'shadow price', a marginal evaluation of that resource, since $\lambda^*(k)$ is the rate of change of the maximum value of output with respect to a change in the availability of that resource. This is a very widespread use of Lagrangian analysis in economics. For example, the fundamental model from which we derive the cost curves for a firm is, *minimize* $C = \Sigma w_i x_i$ subject to $f(x) = y$, where y is a parameter. Using (17), the Lagrange multiplier in this model is the marginal cost function $\partial C^*/\partial y = \lambda^*(w, y)$.

To further show the powerful nature of this analysis, consider the two-factor, two-goods model that plays an important part of international trade theory:

$$\begin{aligned} &\text{maximize } NNP = p_1 y_1 + p_2 y_2 \\ &\text{subject to} \ : \\ &y_1 = f^1(L_1, K_1) \; y_2 = f^2(L_2, K_2) \; L_1 + L_2 \\ &\quad = L \; K_1 + K_2 = K \end{aligned}$$

where f^1 and f^2 are production functions using labour (L) and capital (K) in each of two industries with outputs y_1 and y_2; output prices p_1 and p_2 and labour and capital endowments L and K are parametric. We can enumerate the salient properties of this model just by inspection, using the above results. The Lagrangian for this model is $L = p_1 y_1 + p_2 y_2 + \lambda_1(f^1(L_1, K_1) - y_1) + \lambda_2(f^2(L_2, K_2) - y_2) + \lambda_L(L - L_1 - L_2) + \lambda_K(K - K_1 - K_2)$. Assuming the NFOC and SSOC hold, we solve the NFOC for the output supply functions $y_1^*(p_1, p_2, L, K)$ and $y_2^*(p_1, p_2, L, K)$, and the Lagrange multipliers, particularly $\lambda_L^*(p_1, p_2, L, K)$ and $\lambda_K^*(p_1, p_2, L, K)$. Substituting $y_1^*(\cdot)$ and $y_2^*(\cdot)$ into the objective function, we get the maximum value of *NNP* for given prices and resource constraints, $\phi(p_1, p_2, L, K)$. Since prices enter the objective function only, and in the classic linear form, (9) immediately yields the envelope relations $\phi_{pi} = y_i^*(\cdot)$. We also note $\phi_L = \lambda_L^*(\cdot)$ *and* $\phi_K = \lambda_K^*(\cdot)$. The primal–dual model is, *maximize* $F = p_1 y_1 + p_2 y_2 - \phi(p_1, p_2, L, K)$ subject to the same constraints above. Since p_1 and p_2 do not enter the constraints, F is concave in p_1 and p_2. Since the first two terms are linear and ϕ enters negatively, ϕ is convex in p_1 and p_2, and thus $\phi_{p_i p_i} = \partial y_i^*/p_i > 0$; the supply curves are upward

E

sloping. Furthermore, from (17), the Lagrange multipliers λ_L^* and λ_K^* are the imputed values of labour and capital. If an additional increment of labour, say, became available, λ_L^* would represent its marginal value product, and hence its implied wage in a competitive economy. Without further assumptions (for example, concavity of the production functions), we cannot determine a sign for how these imputed values change when the resource endowment changes: $\partial \lambda_L^* / \partial L / 0$. The reciprocity relationships are straightforward: $\phi_{p_1 p_2} = \partial \lambda_1^* / \partial p_2 = \partial \lambda_2^* / \partial p_1 = \phi_{p_2 p_1}$, and similarly, $\phi_{LK} = \partial \lambda_L^* / \partial K = \partial \lambda_K^* / \partial L = \phi_{KL}$. We also find $\phi_{p_1 L} = \partial \lambda_1^* / \partial L = \partial \lambda_L / \partial p_1 = \phi_{L p_1}$, and so on. It seems unlikely that Jacob Viner could have imagined what the corrected version of his diagram would eventually lead to!

See Also

- ▶ Cost Functions
- ▶ Duality
- ▶ Hicksian and Marshallian Demands
- ▶ Le Chatelier Principle

Bibliography

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Silberberg, E., and W. Suen. 2000. *The structure of economics*. 3rd ed. New York: McGraw-Hill.

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3, 23–46. Repr. in American Economic Association. 1952. *Readings in price theory*. Homewood: Irwin.

Environmental Economics

Robert N. Stavins

Abstract

An overview is provided of the economics of environmental policy, including the setting of goals and targets, notably the Kaldor–Hicks criterion and the related method of assessment known as benefit–cost analysis. Also reviewed are the means of environmental policy, that is, the choice of specific policy instruments, featuring an examination of potential criteria for assessing alternative instruments, with focus on cost-effectiveness. The theoretical foundations and experiential highlights of individual instruments are reviewed, including conventional command-and-control mechanisms and market-based instruments.

Keywords

Asymmetric information; Averting behaviour; Benefit–cost analysis; Bequest value; Coase, R.; Command-and-control instruments; Contingent valuation (CV); Cost- effectiveness; Efficiency; Environmental economics; Existence value; General equilibrium analysis; Hedonic pricing methods; Hedonic wage method; Insurance premium taxes; Kaldor–Hicks criterion; Market-based instruments; Net present value (NPV) analysis; Non-use value; Pareto, V.; Partial equilibrium analysis; Pigou, A.; Pigouvian tax; Pollution charges; Random utility models; Recreation; Reservation price; Revealed preference; Revenue cycling; Risk reduction; Social discount rate; Tax differentiation; Tradable permits; Travel cost method; Use value; Value of a statistical life (VSL); Willingness to accept; Willingness to pay

JEL Classifications

Q5

The fundamental theoretical argument for government activity in the environmental realm is that pollution is an externality – an unintended consequence of market decisions which affect individuals other than the decision maker. Providing incentives for private actors to internalize the full costs of their actions was long thought to be the theoretical solution to the externality problem. The primary advocate of this view was Arthur Pigou, who in *The Economics of Welfare* (1920) proposed that the government should impose a tax

on emissions equal to the cost of the related damages at the efficient level of control.

A response to the Pigouvian perspective was provided by Ronald Coase in 'The problem of social cost' (1960). Coase demonstrated that, in a bilateral bargaining environment with no transaction costs, wealth or income effects, or third-party impacts, two negotiating parties will reach socially desirable agreements, and the overall amount of pollution will be independent of the assignment of property rights. At least some of the specified conditions are unlikely to hold for most environmental problems. Hence, private negotiation will not – in general – fully internalize environmental externalities.

Criteria for Environmental Policy Evaluation

More than 100 years ago Vilfredo Pareto (1896) enunciated the well-known normative criterion for judging whether a social change makes the world better off: a change is *Pareto efficient* if at least one person is made better off and no one is made worse off. This criterion has considerable normative appeal, but virtually no public policies meet the test. Nearly 50 years later Nicholas Kaldor (1939) and John Hicks (1939) postulated a more pragmatic criterion that seeks to identify 'potential Pareto improvements': a change is welfare-improving if those who gain from the change could – in principle – fully compensate the losers, with (at least) one gainer still being better off.

The Kaldor–Hicks criterion – a test of whether total social benefits exceed total social costs – is the theoretical foundation for the use of the analytical device known as benefit–cost (or net present value) analysis. If the objective is to maximize the difference between benefits and costs (net benefits), then the related level of environmental protection (pollution abatement) is defined as the efficient level of protection:

$$\max_{\{q_i\}} \sum_{i=1}^{N} [B_i(q_i) - C_i(q_i)] \rightarrow q_i^* \qquad (1)$$

where q_i is abatement by source i ($i = 1$ to N), $B_i(\cdot)$ is the benefit function for source i, $C_i(\cdot)$ is the cost function for the source, and q_i^* is the efficient level of protection (pollution abatement). The key necessary condition that emerges from the maximization problem of Eq. 1 is that marginal benefits be equated with marginal costs (on the assumption of convexity of the respective functions).

The Kaldor–Hicks criterion is clearly more practical than the strict Pareto criterion, but its normative standing is less solid. Some have argued that other factors should be considered in a measure of social well-being, and that criteria such as distributional equity should trump efficiency considerations in some collective decisions (Sagoff 1993). Many economists would agree with this assertion, and some have noted that the Kaldor–Hicks criterion should be considered neither a necessary nor a sufficient condition for public policy (Arrow et al. 1996).

Benefit–Cost Analysis of Environmental Regulations

The soundness of empirical benefit–cost analysis rests upon the availability of reliable estimates of social benefits and costs, including estimates of the social discount rate. The present value of net benefits (PVNB) is defined as:

$$PVNB = \sum_{t=0}^{T} \left\{ (B_t - C_t) \cdot (1 + \mathrm{r})^{-t} \right\} \qquad (2)$$

where B_t are benefits at time t, C_t are costs at time t, r is the discount rate, and T is the terminal year of the analysis. A positive PVNB means that the policy or project has the potential to yield a Pareto improvement (meets the Kaldor–Hicks criterion). Thus, carrying out benefit–cost or 'net present value' (NPV) analysis requires discounting to translate future impacts into equivalent values that can be compared. In essence, the Kaldor–Hicks criterion provides the rationale both for benefit–cost analysis and for discounting (Goulder and Stavins 2002).

Choosing the discount rate to be employed in an analysis can be difficult, particularly where impacts are spread across a large number of years involving more than a single generation. In

theory, the social discount rate could be derived by aggregating the individual time preference rates of all parties affected by a policy. Evidence from market behaviour and from experimental economics indicates that individuals may employ lower discount rates for impacts of larger magnitude, higher discount rates for gains than for losses, and rates that decline with the time span being considered (Cropper et al. 1994; Cropper and Laibson 1999). In particular, there has been support for the use of hyperbolic discounting and similar approaches with declining discount rates over time (Ainslie 1991; Weitzman 1994, 1998), but most of these approaches are subject to time inconsistency.

The costs of Environmental Regulations

In the environment context, the economist's notion of cost (or, more precisely, opportunity cost) is a measure of the value of whatever must be sacrificed to prevent or reduce the risk of an environmental impact. A full taxonomy of environmental costs ranges from the most obvious to the least direct (Jaffe et al. 1995).

Methods of direct compliance cost estimation, which measure the costs to firms of purchasing and maintaining pollution-abatement equipment plus costs to government of administering a policy, are acceptable when behavioural responses, transitional costs, and indirect costs are small. Partial and general equilibrium analysis allows for the incorporation of behavioural responses to changes in public policy. Partial equilibrium analysis of compliance costs incorporates behavioural responses by modelling supply and/or demand in major affected markets, but assumes that the effects of a regulation are confined to one or a few markets. This may be satisfactory if the markets affected by the policy are small in relation to the overall economy; but, if an environmental policy is expected to have large consequences for the economy, general equilibrium analysis is required, such as through the use of computable general equilibrium models (Hazilla and Kopp 1990; Conrad 2002). The potential interaction of abatement costs with pre-existing taxes indicates the importance of employing general equilibrium models for comprehensive cost analysis. Revenue recycling (using emission tax or auctioned permit revenues to reduce distortionary taxes) can make the costs of pollution control significantly less than they would otherwise be (Goulder 1995).

In a retrospective examination of 28 environmental and occupational safety regulations, Harrington et al. (2000) found that 14 cost estimation analyses had produced ex ante cost estimates that exceeded actual ex post costs, apparently due to technological innovation stimulated by market-based instruments (see below).

The Benefits of Environmental Regulations

Protecting the environment usually involves active employment of capital, labour, and other scarce resources. The benefits of an environmental policy are defined as the sum of individuals' aggregate willingness to pay (WTP) for the reduction or prevention of environmental damages or individuals' willingness to accept (WTA) compensation to tolerate such environmental damages. In theory, which measure of value is appropriate for assessing a particular policy depends upon the related assignment of property rights, the nature of the status quo, and whether the change being measured is a gain or a loss; but under a variety of conditions the difference between the two measures may be expected to be relatively small (Willig 1976). Empirical evidence suggests larger than expected differences between willingness to pay and willingness to accept (Fisher et al. 1988). Theoretical explanations include psychological aversion to loss and poor substitutes for environmental amenities (Hanemann 1991).

The benefits people derive from environmental protection can be categorized as (*a*) related to human health (mortality and morbidity), (*b*) ecological (both market and non-market), or (*c*) materials damage. The distinction between use value and non-use value is critical. In addition to the direct benefits (use value) people receive through protection of their health or through use of a natural resource, they derive passive or non-use value from environmental quality, particularly in the ecological domain. For example, an individual may value a change in an environmental good because she wants to preserve the good for her

heirs (bequest value). Still other people may envision no current or future use by themselves or their heirs, but still wish to protect the good because they believe it should be protected or because they derive satisfaction from simply knowing it exists (existence value).

How much would individuals sacrifice to achieve a small reduction in the probability of death during a given period of time? How much compensation would individuals require to accept a small increase in that probability? These are reasonable economic questions because most environmental regulations result in very small changes in individuals' mortality risks. Hedonic wage studies, averted behaviour, and contingent valuation (all discussed below) can provide estimates of marginal willingness to pay or willingness to accept related to small changes in mortality risk, and such estimates can be normalized as the 'value of a statistical life' (VSL).

The VSL is *not* the value of an individual life, whether in ethical or technical, economic terms. Rather it is simply a convention:

$$VSL = \frac{MWTP \text{ or } MWTA \text{ (from hedonic wage or } CV)}{Small \text{ risk change}} \tag{3}$$

where *MWTP* and *MWTA*, respectively, refer to marginal willingness to pay and marginal willingness to accept. For example, if people are willing, on average, to pay \$12 for a risk reduction from 5 in 500,000 to 4 in 500,000, Eq. 3 would yield:

$$VSL = \frac{\$12}{0.000002} = \$6,000,000 \tag{4}$$

Thus, VSL quantifies the aggregate amount that a group of individuals are willing to pay for small reductions in risk, standardized (extrapolated) for a risk change of 1.0. It is not the economic value of an individual life because the VSL calculation does not signify that an individual would pay \$6 million to avoid (certain) death this year, or accept (certain) death this year in exchange for \$6 million.

Revealed Preference Methods of Environmental Benefit Estimation

The *averting behaviour method*, in which values of willingness to pay are inferred from observations of people's behavioural responses to changes in environmental quality, is grounded in the household production function framework (Bockstael and McConnell 1983). People sometimes take actions to reduce the risk (averting behaviour) or lessen the impacts (mitigating behaviour) of environmental damages, for example by purchasing water filters or bottled water. In theory, people's perceptions of the cost of averting behaviour and its effectiveness should be measured (Cropper and Freeman 1991), but in practice actual expenditures on averting and mitigating behaviours are typically employed. An additional challenge is posed by the necessity of disentangling attributes of the market good or service.

Recreational activities represent a potentially large class of benefits that are important in assessing policies affecting the use of public lands. The models used to estimate recreation demand fall within the class of household production models. *Travel cost models* (or Hotelling–Clawson–Knetsch models) use information about time and money spent visiting a site to infer the value of that recreational resource (Bockstael 1996). The simplest version of the method involves one site and uses data from surveys of users from various geographic origins, together with estimates of the cost of travel and opportunity cost of time, to infer a demand function relating the number of trips to the site to a function of people's willingness to pay for the experience. *Random utility models* explicitly model the consumer's decision to choose a particular site from among recreation locations, assessing the probability of visiting each location. Such models can be used to value changes in environmental quality by comparing decisions to visit alternative sites (Phaneuf and Smith 2004).

All recreation demand models share limitations. First, the valuation of costs depends on estimates of the opportunity cost of (leisure) time, which is notoriously difficult to estimate. Also, most trips to a recreation site are part of a

multi-purpose experience. In addition, random utility models rely on people's perceptions of environmental quality changes. Finally, like all revealed-preference approaches, recreation demand models can be used to estimate use value only; non-use value cannot be examined.

An alternative approach to assessing people's willingness to pay for recreational experiences is to draw on evidence from *private options to use public goods*. This approach also fits within the household production framework, and is based upon the notion of estimating the derived demand for a privately traded option to utilize a freely available public good. In particular, the demand for state fishing licences has been used to infer the benefits of recreational fishing. Using panel data on fishing license sales and prices, combined with data on substitute prices and demographic variables, Bennear et al. (2005) estimated a licence demand function from which the expected benefits of a recreational fishing day were derived.

Hedonic pricing methods are founded on the proposition that people value goods in terms of the bundles of attributes that constitute those goods. *Hedonic property value methods* employ data on residential property values and home characteristics, including structural, neighbourhood, and environmental quality attributes (Palmquist 2003). By regressing the property value on key attributes, the hedonic price function is estimated:

$$P = f\left(\underset{\sim}{x}, \underset{\sim}{z}, e\right) \tag{5}$$

where P = housing price (includes land); $\underset{\sim}{x}$ = vector of structural attributes; $\underset{\sim}{z}$ = vector of neighbourhood attributes; and e = environmental attribute of concern.

From the estimated hedonic price function of Eq. 5, the marginal implicit price of any attribute, including environmental quality, can be calculated as the partial derivative of the housing price with respect to the given attribute:

$$\frac{\partial P}{\partial e} = \frac{\partial f(\cdot)}{\partial e} = P_e \tag{6}$$

This marginal implicit price, P_e, measures the aggregate marginal willingness to pay for the attribute in question. For purposes of benefit estimation, the demand function for the attribute is required, and so it is necessary to examine how the marginal implicit price of the environmental attribute varies with changes in the quantity of the attribute and other relevant variables. If the hedonic price Eq. 5 is nonlinear, then fitted values of P_e can be calculated as e is varied, and a second- stage equation can be estimated:

$$\widehat{P}_e = g\left(e, \underset{\sim}{y}\right) \tag{7}$$

where $\widehat{P}_e$ = the fitted value of the marginal implicit price of e from the first-stage equation; and $\underset{\sim}{y}$ = a vector of factors that affect marginal willingness to pay for e, including buyer characteristics.

Equation 7, above, has been interpreted as the demand function for the environmental attribute, from which benefits (consumers surplus) can be estimated in the usual way; but there are problems. Most important among these is the question of whether a demand function has actually been estimated, since environmental quality may affect both the demand for housing and its supply, raising the classic identification problem. In addition, informational asymmetries may distort the analysis. Also, because the hedonic property method is based on analysis of marginal changes, it should not be applied to analysis of policies with large anticipated effects.

A related benefit-estimation technique is the *hedonic wage method*, based on the reality that individuals in well-functioning labour markets make trade-offs between wages and risk of on-the-job injuries (or death). A job is a bundle of characteristics, including its wage, responsibilities and risk, among others factors. Two jobs that require the same skill level but have different risks of on-the-job mortality will pay different wages. On the labour supply side, employees tend to require extra compensation to accept jobs with greater risks; and on the labour demand side, employers are willing to offer higher wages to attract workers to riskier jobs. Hence, labour market data on wages and job characteristics can be used to estimate people's marginal implicit price

of risk, that is, their valuation of risk. By regressing the wage on key attributes, the hedonic price function is estimated:

$$W = h\left(\underline{x}, r\right) \tag{8}$$

where $W =$ wage (in annual terms); $\underline{x} =$ vector of worker and job characteristics; and $r =$ mortality risk of job.

The marginal implicit price of risk is calculated as the partial derivative of the annual wage with respect to the measured mortality risk:

$$\frac{\partial W}{\partial r} = \frac{\partial h(\cdot)}{\partial r} = W_r \tag{9}$$

This marginal implicit price of risk is the average annual income necessary to compensate a worker for a marginal change in risk throughout the year, and it varies with the level of risk.

Many of the issues that arise with the hedonic property value method have parallels here. First, there is the possibility of simultaneity: causality between risk and wages can run in both directions. Also, if individuals' perceptions of risk do not correspond with actual risks, then the marginal implicit price of risk calculated from a hedonic wage study will be biased, and imperfections in labour markets (less than perfect mobility) can cause further problems.

Direct application of the method in the environmental realm is limited to occupational (as opposed to environmental) exposures and risks. Yet hedonic wage methods are of considerable importance in the environmental policy realm, because the results from hedonic wage studies have frequently been used through 'benefit transfer' to infer the VSL. In such applications, the hedonic wage method brings with it possible bias, because studies typically focus on risky occupations, which may attract workers who are systematically less risk-averse.

Standard economic theory would suggest that younger people would have higher values for risk reduction because they have a longer expected life remaining before them and thus a higher expected lifetime utility (Moore and Viscusi 1988; Cropper and Sussman 1990). In contrast, some models and empirical evidence suggest that older people may in fact have a higher demand for reducing mortality risks than younger people, and that the value of a life may follow an 'inverted-U' shape over the life cycle, with its peak during mid-life (Shepard and Zeckhauser 1982; Mrozek and Taylor 2002; Viscusi and Aldy 2003; Alberini et al. 2004).

E

Stated Preference Methods of Environmental Benefit Estimation

In the best known stated preference method, *contingent valuation* (CV), survey respondents are presented with scenarios that require them to trade off, hypothetically, something for a change in an environmental good or service (Mitchell and Carson 1989; Boyle 2003). The simplest approach is to ask people for their maximum willingness to pay, but as there are few real markets in which individuals are actually asked to generate their reservation prices, this method is considered unreliable. In a bidding game, the researcher begins by stating a willingness-to-pay number, asks for a yes–no response, and then increases or decreases the amount until indifference is achieved. The problem with this approach is starting-point bias. A related approach is the use of a payment card shown to the respondent, but the range of WTP on the card may introduce bias, and the approach cannot be used with telephone surveys. Finally, the referendum (discrete choice) approach is favoured by researchers. Each respondent is offered a different WTP number, to which a simple yes–no response is solicited.

The primary advantage of contingent valuation is that it can be applied to a wide range of situations, including use as well as non-use value; but potential problems remain. Respondents may not understand what they are being asked to value. This may introduce greater variance, if not bias, in responses. Likewise, respondents may not take the hypothetical market seriously because no budget constraint is imposed. This can increase variance and bias. Yet if the scenario is 'too realistic,' strategic bias may be expected to show up in responses. Finally, the 'warm glow effect' may plague some stated preference surveys: people

may purchase moral satisfaction with large but unreal statements of their willingness-to-pay (Andreoni 1995).

The 1989 Exxon Valdez oil spill off the coast of Alaska led to massive litigation, and resulted in the most prominent use ever of the concept of non-use value and the method of contingent valuation for its estimation. The result was a symposium sponsored by the Exxon Corporation attacking the CV method (Hausman 1993), and the subsequent creation of a government panel – established by the National Oceanic and Atmospheric Administration (NOAA) and chaired by two Nobel laureates in economics – to assess the scientific validity of the CV method. The NOAA panel concluded that 'CV studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive (non-use) values' (Arrow et al. 1993, p. 4610). The panel offered its approval of CV methods subject to a set of best-practice guidelines.

It is important to distinguish between legitimate methods of benefit estimation and approaches sometimes encountered in the policy process that do not measure willingness-to-pay or willingness-to-accept. Frequently misused techniques include: (*a*) employing, as proxies for the benefits of a policy, estimates of the 'cost avoided' by not using the next most costly means of achieving the policy's goals; (*b*) 'societal revealed preference' models, which seek to infer the benefits of a proposed policy from the costs of previous regulatory actions; and (*c*) cost-of-illness or human-capital measures which estimate explicit market costs resulting from changes in morbidity or mortality. Because none of these approaches provides estimates of WTP or WTA, these techniques do not provide valid measures of economic benefits.

Choosing Instruments: The Means of Environmental Policy

Even if the goals of environmental policies are given, economic analysis can bring insights to the assessment and design of environmental policies. One important criterion is *cost-effectiveness*, defined as the allocation of control among sources that results in the aggregate target being achieved at the lowest possible cost, that is, the allocation which satisfies the following cost-minimization problem:

$$\min_{\{r_i\}} C = \sum_{i=1}^{N} c_i(r_i) \tag{10}$$

$$s.t. \sum_{i=1}^{N} [u_i - r_i] \leq E \tag{11}$$

$$and\ 0 \leq r_i \leq u_i \tag{12}$$

where r_i = reductions in emissions (abatement or control) by source i ($i = 1$ to N); $c_i(r_i)$ = cost function for source i; C = aggregate cost of control; u_i = uncontrolled emissions by source i; and E = the aggregate emissions target imposed by the regulatory authority.

If the cost functions are convex, then necessary and sufficient conditions for satisfaction of the constrained optimization problem posed by Eqs. 10–12 are the following (among others) (Kuhn and Tucker 1951):

$$\frac{\partial c_i(r_i)}{\partial r_i} - \lambda \geq 0 \tag{13}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - \lambda\right] = 0 \tag{14}$$

Equations 13 and 14 together imply the crucial condition for cost- effectiveness that all sources (that exercise some degree of control) experience the same marginal abatement costs (Baumol and Oates 1988). Thus, when one examines environmental policy instruments, a key question is whether marginal abatement costs are likely to be being equated across sources.

Command-and-Control Versus Market-Based Instruments

Conventional approaches to regulating the environment – frequently characterized as command-and-control – allow relatively little flexibility in the means of achieving goals. Such

policy instruments tend to force firms to take on equal shares of the pollution-control burden, regardless of the cost. The most prevalent form of uniform command-and-control standards is technology standards that specify the adoption of specific pollution-control technologies, and performance standards that specify uniform limits on the amount of pollution a facility can generate. In theory, non- uniform performance standards could be made to be cost-effective, but the government typically lacks the requisite information (on marginal costs of individual sources).

Market-based instruments encourage behaviour through market signals rather than through explicit directives regarding pollution-control levels or methods. Market-based instruments fall within four categories: pollution charges, tradable permits, market-friction reductions, and government subsidy reductions. Liability rules may also be thought of as a market-based instrument, because they provide incentives for firms to take into account the potential environmental damages of their decisions.

Where there is significant heterogeneity of abatement costs, command-and- control methods will not be cost-effective. In reality, costs can vary enormously due to production design, physical configuration, age of assets, and other factors. For example, the marginal costs of controlling lead emissions have been estimated to range from $13 to $56,000 per ton (Hartman et al. 1994; Morgenstern 2000). But where costs are similar among sources, command-and- control instruments may perform as well as (or better than) market-based instruments, depending on transactions costs, administrative costs, possibilities for strategic behaviour, political costs, and the nature of the pollutants (Newell and Stavins 2003).

In theory, market-based instruments allow any desired level of pollution clean- up to be realized at the lowest overall cost by providing incentives for the greatest reductions in pollution by those firms that can achieve the reductions most cheaply. Rather than equalizing pollution levels among firms, market-based instruments equalize their marginal abatement costs (Montgomery 1972). In addition, market- based instruments have the potential to bring down abatement costs over time by providing incentives for companies to adopt cheaper and better pollution-control technologies. This is because, with market-based instruments, most clearly with emission taxes, it pays firms to clean up a bit more if a sufficiently low-cost method (technology or process) of doing so can be identified and adopted (Downing and White 1986; Maleug 1989; Milliman and Prince 1989; Jaffe and Stavins 1995). However, the ranking among policy instruments in terms of their respective impacts on technology innovation and diffusion is ambiguous (Jaffe et al. 2003).

Closely related to the effects of instrument choice on technological change are the effects of vintage-differentiated regulation on the rate of capital turnover, and thereby on pollution abatement costs and environmental performance. Vintage- differentiated regulation is a common feature of many environmental policies, whereby the standard for regulated units is fixed in terms of their date of entry, with later vintages facing more stringent regulation. Such vintage-differentiated regulations can be expected to retard turnover in the capital stock, and thereby to reduce the cost-effectiveness of regulation. Under some conditions the result can be higher levels of pollutant emissions than would occur in the absence of regulation. Such economic and environmental consequences are not only predictions from theory (Maloney and Brady 1988); both types of consequences have been validated empirically (Gruenspecht 1982; Nelson et al. 1993).

Pollution Charges

Pollution charge systems assess a fee or tax on the amount of pollution that firms or sources generate (Pigou 1920). By definition, actual emissions are equal to unconstrained emissions minus emissions reductions, that is, $e_i = u_i - r_i$. A source's cost minimization problem in the presence of an emissions tax, t, is given by:

$$\min_{\{r_i\}} \; [c_i(r_i) + t \cdot (u_i - r_i)] \tag{15}$$

$$s.t. \;\; r_i \geq 0 \tag{16}$$

The result for each source is:

$$\frac{\partial c_i(r_i)}{\partial r_i} - t \geq 0 \tag{17}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - t\right] = 0 \tag{18}$$

Equations 17 and 18 imply that each source (that exercises a positive level of control) will carry out abatement up to the point where its marginal control costs are equal to the tax rate. Hence, marginal abatement costs are equated across sources, satisfying the condition for cost-effectiveness specified by Eqs. 13 and 14, at least in the simplest case of a uniformly mixed pollutant. In the non-uniformly mixed pollutant case, where 'hot spots' can be an issue, the respective cost-effective instrument is an 'ambient charge'.

A challenge with charge systems is identifying the appropriate tax rate. For social efficiency, it should be set equal to the marginal benefits of clean-up at the efficient level of clean-up (Pigou 1920); but policymakers are more likely to think in terms of a desired level of clean-up, and they do not know beforehand how firms will respond to a given level of taxation. An additional problem is that, although such systems minimize aggregate social costs, these systems may be *more* costly than comparable command-and-control instruments *for regulated firms*, because firms pay both their abatement costs *and* taxes on their residual emissions.

If charges are broadly defined, many applications can be identified (Stavins 2003). Coming closest to true Pigouvian taxes are the increasingly common *unit- charge* systems for financing municipal solid waste collection, where households and businesses are charged the incremental costs of collection and disposal. Another important set of charge systems has been *deposit refund systems*, whereby consumers pay a surcharge when purchasing potentially polluting products, and receive a refund when returning the product to an approved centre for recycling or disposal. A number of countries and states have implemented this approach to control litter from beverage containers and to reduce the flow of solid waste to landfills (Bohm 1981; Menell 1990), and the concept has also been applied to lead-acid batteries. There has also been considerable use of *environmental user charges*, through which specific environmentally related services are funded. Examples include *insurance premium taxes* (Barthold 1994). Another set of environmental charges are *sales taxes* on motor fuels, ozone-depleting chemicals, agricultural inputs, and low-mileage motor vehicles. Finally, *tax differentiation* has been used to encourage the use of renewable energy sources.

Tradable Permit Systems

Tradable permits can achieve the same cost-minimizing allocation as a charge system, while avoiding the problems of uncertain firm responses and the distributional consequences of taxes. Under a tradable permit system, an allowed overall level of pollution, E^{-}, is established, and allocated among sources in the form of permits. Firms that keep emission levels below allotted levels may sell surplus permits to other firms or use them to offset excess emissions in other parts of their operations. Let q_{0i} be the initial allocation of emission permits to source i, such that:

$$\sum_{i=1}^{N} q_{0i} = \overline{E} \tag{19}$$

Then, if p is the market-determined price of tradable permits, a single firm's cost minimization problem is given by:

$$\min_{\{r_i\}} \left[c_i(r_i) + p \cdot (u_i - r_i - q_{0i})\right] \tag{20}$$

$$s.t. r_i \geq 0 \tag{21}$$

The result for each source is:

$$\frac{\partial c_i(r_i)}{\partial r_i} - p \geq 0 \tag{22}$$

$$r_i \cdot \left[\frac{\partial c_i(r_i)}{\partial r_i} - p\right] = 0 \tag{23}$$

Equations 22 and 23 together imply that each source (that exercises a positive level of control)

will carry out abatement up to the point where its marginal control costs are equal to the market-determined permit price. Hence, the environmental constraint, E, is satisfied, and marginal abatement costs are equated across sources, satisfying the condition of cost-effectiveness. The unique cost-effective equilibrium is achieved independently of the initial allocation of permits (Montgomery 1972), which is of great political significance.

The performance of a tradable permit system can be adversely affected by: concentration in the permit market (Hahn 1984; Misolek and Elder 1989); concentration in the product market (Maleug 1990); transaction costs (Stavins 1995); non-profit maximizing behaviour, such as sales or staff maximization (Tschirhart 1984); the pre-existing regulatory environment (Bohi and Burtraw 1992); and the degree of monitoring and enforcement (Montero 2003).

Tradable permits have been the most frequently used market-based system (US Environmental Protection Agency 2000). Significant applications include: the emissions trading programme (Tietenberg 1985; Hahn 1989); the leaded gasoline phase-down; water quality permit trading (Hahn 1989; Stephenson et al. 1998); CFC trading (Hahn and McGartland 1989); the sulphur dioxide (SO2) allowance trading system for acid rain control (Schmalensee et al. 1998; Stavins 1998; Carlson et al. 2000; Ellerman et al. 2000); the RECLAIM programme in the Los Angeles metropolitan region (Harrison 1999); tradable development rights for land use; and the European Union's greenhouse gas emission trading scheme.

Market Friction Reduction

Market friction reduction can serve as a policy instrument for environmental protection. *Market creation* establishes markets for inputs or outputs associated with environmental quality. Examples of market creation include measures that facilitate the voluntary exchange of water rights and thus promote more efficient allocation and use of scarce water supplies (Howe 1997), and policies that facilitate the restructuring of electricity generation and transmission. Since well-functioning markets depend, in part, on the existence of well-informed producers and consumers, *information programmes* can help foster market-oriented solutions to environmental problems. These programmes have been of two types. *Product labelling requirements* have been implemented to improve information sets available to consumers, while other programmes have involved *reporting requirements* (Hamilton 1995; Konar and Cohen 1997; Khanna et al. 1998).

Government Subsidy Reduction

Government subsidy reduction constitutes another category of market-based instruments. Subsidies are the mirror image of taxes and, in theory, can provide incentives to address environmental problems. Although subsidies can advance environmental quality (see, for example, Jaffe and Stavins 1995), it is also true that subsidies, in general, have important disadvantages relatives to taxes (Dewees and Sims 1976; Baumol and Oates 1988). Because subsidies increase profits in an industry, they encourage entry, and can thereby increase industry size and pollution output (Mestelman 1982; Kohn 1985). In practice, rather than internalizing externalities, many subsidies promote economically inefficient and environmentally unsound practices. In such cases, reducing subsidies can increase efficiency and improve environmental quality. For example, because of concerns about global climate change, increased attention has been given to cutting inefficient subsidies that promote the use of fossil fuels.

Implications of Uncertainty for Instrument Choice

The dual task facing policymakers of choosing environmental goals and selecting policy instruments to achieve those goals must be carried out in the presence of the significant uncertainty that affects the benefits and the costs of environmental protection. Since Weitzman's (1974) classic paper on 'Prices vs. quantities', it has been widely acknowledged that benefit uncertainty on its own has no effect on the identity of the efficient control instrument, but that cost uncertainty can have significant effects, depending upon the relative

slopes of the marginal benefit (damage) and marginal cost functions. In particular, if uncertainty about marginal abatement costs is significant, and if marginal abatement costs are flat relative to marginal benefits, then a quantity instrument is more efficient than a price instrument.

In the environmental realm, benefit uncertainty and cost uncertainty are usually both present, with benefit uncertainty of greater magnitude. When marginal benefits are positively correlated with marginal costs (which, it turns out, is not uncommon), then there is an additional argument in favour of the relative efficiency of quantity instruments (Stavins 1996). Nevertheless, the regulation of stock pollutants will often favour price instruments, because the marginal benefit function – linked with the stock of pollution – will tend to be flatter than the marginal cost function – linked with the flow of pollution (Newell and Pizer 2003). In theory, there would be considerable efficiency advantages in the presence of uncertainty of hybrid systems – for example, quotas combined with taxes – or nonlinear taxes (Roberts and Spence 1976; Weitzman 1978; Kaplow and Shavell 2002; Pizer 2002), but such systems have not been adopted.

Conclusion

The growing use of economic analysis to inform environmental decision-making marks greater acceptance of the usefulness of these tools in improving regulation. But debates about the normative standing of the Kaldor–Hicks criterion and the challenges inherent in making benefit–cost analysis operational will continue. Nevertheless, economic analysis has assumed a significant position in the regulatory state. At the same time, despite the arguments made for decades by economists, there is only limited political support for broader use of benefit–cost analysis to assess proposed or existing environmental regulations. These analytical methods remain on the periphery of policy formulation. In a growing literature (not reviewed here), economists have examined the processes through which political decisions regarding environmental regulation are made (Stavins 2004).

The significant changes that have taken place over the past 20 years with regard to the means of environmental policy – that is, acceptance of market-based environmental instruments – may provide a model for progress with analysis of the ends – the targets and goals – of public policies in this domain. The change in the former realm has been dramatic. Market-based instruments have moved centre stage, and policy debates today look very different from those of 20 years ago, when these ideas were routinely characterized as 'licences to pollute' or dismissed as completely impractical. Market-based instruments are now considered seriously for nearly every environmental problem that is tackled, ranging from endangered species preservation to regional smog and global climate change. Of course, no individual policy instrument – whether market-based or conventional – is appropriate for all environmental problems. Which instrument is best in any given situation depends upon a variety of characteristics of the environmental problem, and the social, political, and economic context in which it is regulated.

See Also

- ▶ Climate Change, Economics of
- ▶ Common Property Resources
- ▶ Contingent Valuation
- ▶ Ecological Economics
- ▶ Energy Economics
- ▶ Environmental Kuznets Curve
- ▶ Hedonic Prices
- ▶ Household Production and Public Goods
- ▶ Pollution Haven Hypothesis
- ▶ Pollution Permits
- ▶ Social Discount Rate
- ▶ Value of Life

Bibliography

Ainslie, G. 1991. Derivation of rational economic behavior from hyperbolic discount curves. *American Economic Review* 81: 334–340.

Alberini, A., M. Cropper, A. Krupnick, and N. Simon. 2004. Does the value of a statistical life vary with age

and health status? Evidence from the US and Canada. *Journal of Environmental Economics and Management* 48: 769–792.

Andreoni, J. 1995. Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110: 1–21.

Arrow, K., R. Solow, P. Portney, E. Leamer, R. Radner, and H. Schuman. 1993. Report of the NOAA panel on contingent valuation. *Federal Register* 58: 4601–4614.

Arrow, K., M. Cropper, G. Eads, R. Hahn, L. Lave, R. Noll, P. Portney, M. Russell, R. Schmalensee, K. Smith, and R. Stavins. 1996. Is there a role for benefit–cost analysis in environmental, health, and safety regulation? *Science* 272: 221–222.

Barthold, T. 1994. Issues in the design of environmental excise taxes. *Journal of Economic Perspectives* 8 (1): 133–151.

Baumol, W., and W. Oates. 1988. *The theory of environmental policy.* Cambridge: Cambridge University Press.

Bennear, L., R. Stavins, and A. Wagner. 2005. Using revealed preferences to infer environmental benefits: Evidence from recreational fishing. *Journal of Regulatory Economics* 28: 157–179.

Bockstael, N. 1996. Travel cost methods. In *The handbook of environmental economics*, ed. D. Bromley. Oxford: Blackwell Publishers.

Bockstael, N., and K. McConnell. 1983. Welfare measurement in the household production framework. *American Economic Review* 73: 806–814.

Bohi, D., and D. Burtraw. 1992. Utility investment behavior and the emission trading market. *Resources and Energy* 14: 129–153.

Bohm, P. 1981. *Deposit-refund systems: Theory and applications to environmental, conservation, and consumer policy.* Baltimore: Resources for the Future/Johns Hopkins University Press.

Boyle, K. 2003. Contingent valuation in practice. In *A primer on nonmarket valuation*, ed. P. Champ, K. Boyle, and T. Brown. Dordrecht: Kluwer Academic Publishers.

Carlson, C., D. Burtraw, M. Cropper, and K. Palmer. 2000. Sulfur dioxide control by electric utilities: What are the gains from trade? *Journal of Political Economy* 108: 1292–1326.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Conrad, K. 2002. Computable general equilibrium models in environmental and resource economics. In *The international yearbook of environmental and resource economics 2002/2003: A survey of current issues*, ed. T. Tietenberg and H. Folmer. Northampton: Edward Elgar.

Cropper, M., and A. Freeman. 1991. Environmental health effects. In *Measuring the demand for environmental quality*, ed. B. Braden and C. Kolstad. Amsterdam: Elsevier Science Publications.

Cropper, M., and D. Laibson. 1999. The implications of hyperbolic discounting for project evaluation. In *Discounting and intergenerational equity*, ed. P. Portney and J. Weyant. Washington, DC: Resources for the Future.

Cropper, M., and F. Sussman. 1990. Valuing future risks to life. *Journal of Environmental Economics and Management* 19: 160–174.

Cropper, M., S. Aydede, and P. Portney. 1994. Preferences for life saving programs: How the public discounts time and age. *Journal of Risk and Uncertainty* 8: 243–265.

Dewees, D., and W. Sims. 1976. The symmetry of effluent charges and subsidies for pollution control. *Canadian Journal of Economics* 9: 323–331.

Downing, P., and L. White. 1986. Innovation in pollution control. *Journal of Environmental Economics and Management* 13: 18–27.

Ellerman, D., P. Joskow, R. Schmalensee, J. Montero, and E. Bailey. 2000. *Markets for clean air: The US acid rain program*. New York: Cambridge University Press.

Fisher, A., G. McClelland, and W. Schulze. 1988. Measures of willingness to pay versus willingness to accept: Evidence, explanations, and potential reconciliation. In *Amenity resource valuation: Integrating economics with other disciplines*, ed. G. Peterson, B. Driver, and R. Gregory. State College: Venture.

Goulder, L. 1995. Environmental taxation and the double dividend: A reader's guide. *International Tax and Public Finance* 2: 157–183.

Goulder, L., and R. Stavins. 2002. An eye on the future: How economists' controversial practice of discounting really affects the evaluation of environmental policies. *Nature* 419: 673–674.

Gruenspecht, H. 1982. Differentiated regulation: The case of auto emissions standards. *American Economic Review: Papers and Proceedings* 72: 328–331.

Hahn, R. 1984. Market power and transferable property rights. *Quarterly Journal of Economics* 99: 753–765.

Hahn, R. 1989. Economic prescriptions for environmental problems: How the patient followed the doctor's orders. *Journal of Economic Perspectives* 3: 95–114.

Hahn, R., and A. McGartland. 1989. Political economy of instrument choice: An examination of the U.S. role in implementing the Montreal Protocol. *Northwestern University Law Review* 83: 592–611.

Hamilton, J.T. 1995. Pollution as news: Media and stock market reactions to the toxic release inventory data. *Journal of Environmental Economics and Management* 28: 98–113.

Hanemann, M. 1991. Willingness to pay and willingness to accept: How much can they differ? *American Economic Review* 81: 635–647.

Harrington, W., R. Morgenstern, and P. Nelson. 2000. On the accuracy of regulatory cost estimates. *Journal of Policy Analysis and Management* 19: 297–322.

Harrison, D. Jr. 1999. Turning theory into practice for emissions trading in the Los Angeles air basin. In *Pollution for sale: Emissions trading and joint implementation*, ed. S. Sorrell and J. Skea. London: Edward Elgar.

Hartman, R., D. Wheeler, and M. Singh. 1994. *The cost of air pollution abatement*, Policy research working paper o. 1398. Washington, DC: World Bank.

Hausman, J., ed. 1993. *Contingent valuation: A critical assessment*. Amsterdam: North-Holland.

Hazilla, M., and R. Kopp. 1990. Social cost of environmental quality regulations: A general equilibrium analysis. *Journal of Political Economy* 98: 853–873.

Hicks, J. 1939. The foundations of welfare economics. *Economic Journal* 49: 696–712.

Howe, C. 1997. Increasing efficiency in water markets: Examples from the Western United States. In *Water marketing – the next generation*, ed. T. Anderson and P. Hill. Lanham: Rowman and Littlefield Publishers.

Jaffe, A., and R. Stavins. 1995. Dynamic incentives of environmental regulation: The effects of alternative policy instruments on technological diffusion. *Journal of Environmental Economics and Management* 29: S43–S63.

Jaffe, A., S. Peterson, P. Portney, and R. Stavins. 1995. Environmental regulation and the competitiveness of U.S. manufacturing: What does the evidence tell us? *Journal of Economic Literature* 33: 132–165.

Jaffe, A., R. Newell, and R. Stavins. 2003. Technological change and the environment. In *The handbook of environmental economics*, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/Elsevier Science.

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49: 549–552.

Kaplow, L., and S. Shavell. 2002. On the superiority of corrective taxes to quantity regulation. *American Law and Economics Review* 4: 1–17.

Khanna, M., W. Quimio, and D. Bojilova. 1998. Toxic release information: A policy tool for environmental protection. *Journal of Environmental Economics and Management* 36: 243–266.

Kohn, R. 1985. A general equilibrium analysis of the optimal number of firms in a polluting industry. *Canadian Journal of Economics* 18: 347–354.

Konar, S., and M. Cohen. 1997. Information as regulation: The effect of community right to know laws on toxic emissions. *Journal of Environmental Economics and Management* 32: 109–124.

Kuhn, H., and A. Tucker. 1951. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Maleug, D. 1989. Emission credit trading and the incentive to adopt new pollution abatement technology. *Journal of Environmental Economics and Management* 16: 52–57.

Maleug, D. 1990. Welfare consequences of emission credit trading programs. *Journal of Environmental Economics and Management* 18: 66–77.

Maloney, M., and G. Brady. 1988. Capital turnover and marketable property rights. *Journal of Law and Economics* 31: 203–226.

Menell, P. 1990. Beyond the throwaway society: An incentive approach to regulating municipal solid waste. *Ecology Law Quarterly* 17: 655–739.

Mestelman, S. 1982. Production externalities and corrective subsidies: A general equilibrium analysis. *Journal of Environmental Economics and Management* 9: 186–193.

Milliman, S., and R. Prince. 1989. Firm incentives to promote technological change in pollution control. *Journal of Environmental Economics and Management* 17: 247–265.

Misolek, W., and H. Elder. 1989. Exclusionary manipulation of markets for pollution rights. *Journal of Environmental Economics and Management* 16: 156–166.

Mitchell, R., and R. Carson. 1989. *Using surveys to value public goods: The contingent valuation method*. Washington, DC: Resources for the Future.

Montero, J. 2003. *Tradeable permits with imperfect monitoring: Theory and evidence*, Working paper. Cambridge, MA: Center for Energy and Environmental Policy, MIT.

Montgomery, D. 1972. Markets in licenses and efficient pollution control programs. *Journal of Economic Theory* 5: 395–418.

Moore, M., and W. Viscusi. 1988. The quantity-adjusted value of life. *Economic Inquiry* 26: 369–388.

Morgenstern, R. 2000. Decision making at EPA: Economics, incentives and efficiency. Draft paper for a conference: EPA at thirty: Evaluating and improving the environmental protection agency. Duke University, 7–8 December.

Mrozek, J., and L. Taylor. 2002. What determines the value of life? A meta analysis. *Journal of Policy Analysis and Management* 21: 253–270.

Nelson, R., T. Tietenberg, and M. Donihue. 1993. Differential environmental regulation: Effects on electric utility capital turnover and emissions. *The Review of Economics and Statistics* 75: 368–373.

Newell, R., and W. Pizer. 2003. Regulating stock externalities under uncertainty. *Journal of Environmental Economics and Management* 45: 416–432.

Newell, R., and R. Stavins. 2003. Cost heterogeneity and the potential savings from market-based policies. *Journal of Regulatory Economics* 23: 43–59.

Palmquist, R. 2003. Property value models. In *Handbook of environmental economics*, ed. K. Mäler and J. Vincent, vol. 2. Amsterdam: North-Holland/Elsevier Science.

Pareto, V. 1896. *Cours d'Economie Politique*, vol. 2. Lausanne: F. Rouge.

Phaneuf, D., and V. Smith. 2004. Recreation demand models. In *Handbook of environmental economics*, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/Elsevier Science.

Pigou, A. 1920. *The economics of welfare*. London: Macmillan.

Pizer, W. 2002. Combining price and quantity controls to mitigate global climate change. *Journal of Public Economics* 85: 409–434.

Porter, M., and C. van der Linde. 1995. Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives* 9 (4): 97–118.
Revesz, R., and R. Stavins. 2005. Environmental law and policy. In *The handbook of law and economics*, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland/Elsevier Science.
Roberts, M., and M. Spence. 1976. Effluent charges and licenses under uncertainty. *Journal of Public Economics* 5: 193–197.
Sagoff, M. 1993. Environmental economics: An epitaph. *Resources* 111: 2–7.
Schmalensee, R., P. Joskow, A. Ellerman, J. Montero, and E. Bailey. 1998. An interim evaluation of sulfur dioxide emissions trading. *Journal of Economic Perspectives* 12 (3): 53–68.
Shepard, D., and R. Zeckhauser. 1982. Life-cycle consumption and willingness to pay for increased survival. In *The value of life and safety*, ed. M. Jones-Lee. Amsterdam: North-Holland.
Stavins, R. 1995. Transaction costs and tradeable permits. *Journal of Environmental Economics and Management* 29: 133–146.
Stavins, R. 1996. Correlated uncertainty and policy instrument choice. *Journal of Environmental Economics and Management* 30: 218–225.
Stavins, R. 1998. What have we learned from the grand policy experiment: Lessons from SO2 allowance trading. *Journal of Economic Perspectives* 12 (3): 69–88.
Stavins, R. 2003. Experience with market-based environmental policy instruments. In *The handbook of environmental economics*, ed. K. Mäler and J. Vincent. Amsterdam: North-Holland/Elsevier Science.
Stavins, R. 2004. *The political economy of environmental regulation*. Northampton: Edward Elgar.
Stephenson, K., P. Norris, and L. Shabman. 1998. Watershed-based effluent trading: The nonpoint source challenge. *Contemporary Economic Policy* 16: 412–421.
Tietenberg, T. 1985. *Emissions trading: An exercise in reforming pollution policy*. Washington, DC: Resources for the Future.
Tschirhart, J. 1984. Transferable discharge permits and the control of stationary source air pollution: A survey and synthesis. In *Economic perspectives on acid deposition control*, ed. T. Crocker. Boston: Butterworth.
US Environmental Protection Agency. 2000. *Guidelines for preparing economic analyses*. Washington, DC: US EPA.
Viscusi, W., and J. Aldy. 2003. The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty* 27 (1): 5–76.
Weitzman, M. 1974. Prices vs. quantities. *Review of Economic Studies* 41: 477–491.
Weitzman, M. 1978. Optimal rewards for economic regulation. *American Economic Review* 68: 683–691.
Weitzman, M. 1994. On the environmental discount rate. *Journal of Environmental Economics and Management* 26: 200–209.
Weitzman, M. 1998. Why the far-distant future should be discounted at its lowest possible rate. *Journal of Environmental Economics and Management* 36: 201–208.
Willig, R. 1976. Consumer's surplus without apology. *American Economic Review* 66: 589–597.

Environmental Kuznets Curve

Arik Levinson

Abstract

Pollution often appears first to worsen and later to improve as countries' incomes grow. Because of its resemblance to the pattern of inequality and income described by Simon Kuznets, this pattern of pollution and income has been labelled an 'environmental Kuznets curve'. While many pollutants exhibit this pattern, peak pollution levels occur at different income levels for different pollutants, countries and time periods. This link between income and pollution cannot be interpreted causally, and is consistent with either efficient or inefficient growth paths. The evidence does, however, refute the claim that environmental degradation is an inevitable consequence of economic growth.

Keywords

Carbon emissions; Economic growth: and pollution; Environmental Kuznets curve; Environmental regulation; Local externalities; Pollution; Income

JEL Classification

Q56

Some forms of pollution appear first to worsen and later to improve as countries' incomes grow. The world's poorest and richest countries have relatively clean environments, while middle-income countries are the most polluted. Because

of its resemblance to the pattern of inequality and income described by Simon Kuznets (1955), this pattern of pollution and income has been labelled an 'environmental Kuznets curve' (EKC).

Grossman and Krueger (1995) and the World Bank (1992) first popularized this idea, using a simple empirical approach. They regress data on ambient air and water quality in cities worldwide on a polynomial in GDP per capita and other city and country characteristics. They then plot the fitted values of pollution levels as a function of GDP per capita, and demonstrate that many of the plots appear inverse-U-shaped, first rising and then falling. The peaks of these predicted pollution-income paths vary across pollutants, but 'in most cases they come before a country reaches a per capita income of $8000' in 1985 dollars (Grossman and Krueger 1995, p. 353).

In the years since these original observations were made, researchers have examined a wide variety of pollutants for evidence of the EKC pattern, including automotive lead emissions, deforestation, greenhouse gas emissions, toxic waste and indoor air pollution. Some investigators have experimented with different econometric approaches, including higher-order polynomials, fixed and random effects, splines, semi- and non-parametric techniques, and different patterns of interactions and exponents. Others have studied different groups of jurisdictions and different time periods, and have added control variables, including measures of corruption, democratic freedoms, international trade openness, and even income inequality (bringing the subject full circle back to Kuznets's original idea).

Some generalizations across these approaches emerge. Roughly speaking, pollution involving local externalities begins improving at the lowest income levels. Fecal coliform in water and indoor household air pollution are examples. For some of these local externalities, pollution appears to decrease steadily with economic growth, and we observe no turning point at all. This is not a rejection of the EKC; pollution must have increased at some point in order to decline with income eventually, and there simply are no data from the earlier period. By contrast, pollutants involving very dispersed externalities tend to have their turning points at the highest incomes, or even no turning points at all, as pollution appears to increase steadily with income. Carbon emissions provide one such example. This, too, is not necessarily a rejection of the EKC; the turning points for these pollutants may come at levels of income per capita higher than in today's wealthiest economies.

Another general empirical result is that the turning points for individual pollutants differ across countries. This difference shows up as instability in empirical approaches that estimate one fixed turning point for any given pollutant. Countries that are the first to deal with a pollutant do so at higher income levels than following countries, perhaps because the following countries benefit from the science and engineering lessons of the early movers.

Most researchers have been careful to avoid interpreting these reduced-form empirical correlations structurally, and to recognize that economic growth does not automatically cause environmental improvements. All of the studies omit country characteristics correlated with both income and pollution levels, the most important being environmental regulatory stringency. The EKC pattern does not provide evidence of market failures or efficient policies in rich or poor countries. Rather, there are multiple underlying mechanisms, some of which have begun to be modelled theoretically.

In theory, the EKC relationship can be divided into three parts: scale, composition, and technique (see Brock and Taylor 2005). If as an economy grows the *scale* of all activities increases proportionally, pollution will increase with economic growth. If growth is not proportional but is accompanied by a change in the *composition* of goods produced, then pollution may decline or increase with income. If richer economies produce proportionally fewer pollution-intensive products, because of changing tastes or patterns of trade, this composition effect can lead to a decline in pollution associated with economic growth. Finally, if richer countries use less pollution-intensive production *techniques*, perhaps because environmental quality is a normal good, growth can lead to falling pollution. The EKC summarizes the interaction of these three processes.

Beyond this aggregate decomposition of the EKC, some attempts have been made to formalize structural models that lead to inverse-U-shaped pollution-income patterns. Many describe economies at some type of corner solution initially, where residents of poor countries are willing to trade environmental quality for income at a faster rate than possible using available technologies or resources. As the model economies become wealthier and their environments dirtier, eventually the marginal utility of income falls and the marginal disutility from pollution rises, to the point where people choose costly abatement mechanisms. After that point, the economies are at interior solutions, marginal abatement costs equal marginal rates of substitution between environmental quality and income, and pollution declines with income (see Stokey 1998). In frameworks of this type, there is typically zero pollution abatement until some threshold income level is crossed, after which abatement begins and pollution starts declining with income.

To date, the practical lessons from this theoretical literature are limited. Most of the models are designed to yield inverse-U-shaped pollution-income paths, and succeed using a variety of assumptions and mechanisms. Hence, any number of forces may be behind the empirical observation that pollution increases and then decreases with income. Moreover, that pattern cannot be interpreted causally, and is consistent with either efficient or inefficient growth paths. Perhaps the most important insight is in Grossman and Krueger's original paper: 'We find no evidence that economic growth does unavoidable harm to the natural habitat' (1995, p. 370). Economists have long argued that environmental degradation is not an inevitable consequence of economic growth. The EKC literature provides empirical support for that claim.

See Also

- ▶ Environmental Economics
- ▶ Growth and International Trade
- ▶ Pollution Haven Hypothesis

Bibliography

Brock, W., and M. Taylor. 2005. Economic growth and the environment: A review of theory and empirics. In *The handbook of economic growth*, vol. 1, ed. S. Durlauf and P. Aghion. Amsterdam: North-Holland.

Grossman, G., and A. Krueger. 1995. Economic growth and the environment. *Quarterly Journal of Economics* 110: 353–377.

Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45: 1–28.

Stokey, N. 1998. Are there limits to growth? *International Economic Review* 39: 1–31.

World Bank. 1992. *World development report 1992*. New York: Oxford University Press.

Envy

Peter J. Hammond

Envy is a deadly sin, but then so is avarice or greed, and greed seems not to trouble economists. Envy does, however, perhaps because it is an externality. Different economists have also used the term in different senses. Veblen (1899) avoids the word 'envy', but one feels that some of the pleasure of conspicuous consumption may come from the malicious belief that it induces envy in others. Brennan (1973) uses the term 'malice' to indicate negative altruism – a distaste for the income of others – and 'envy' to indicate that the marginal disutility of another's income increases as their income increases. For other concepts of envy, see Nozick (1974) and Chaudhuri (1985).

Most economists now use the word 'envy' in a narrow technical sense due to Foley (1967), who was much more interested, however, in finding an adequate concept of 'equity'. First, however, one should turn to Rawls (1971), whose *Theory of Justice* has 12 pages of very pertinent discussion.

Rawls on Envy and Justice

Rawls (1971) defines envy in a way which he attributes to Kant, and he is careful to distinguish

'envy' from 'jealousy', which can be thought of as a protective response to envy:

> we may think of envy as the propensity to view with hostility the greater good of others even though their being more fortunate than we are does not detract from our advantages. We envy persons whose situation is superior to ours... and we are willing to deprive them of their greater benefits even if it is necessary to give up something ourselves. When others are aware of our envy, they may become jealous of their better circumstances and anxious to take precautions against the hostile acts to which our envy makes us prone. So understood envy is collectively disadvantageous: the individual who envies another is prepared to do things that make them both worse off, if only the discrepancy between them is sufficiently reduced (p. 532).

This is in section 80, on 'the problem of envy' in which Rawls asks whether his theory of justice is likely to prove impractical because 'just institutions... are likely to arouse and encourage these propensities [such as envy] to such an extent that the social system becomes unworkable and incompatible with human good' (p. 531). This is a positive question; a normative one arises when one recognizes the possibility of 'excusable envy' because 'sometimes the circumstances evoking envy are so compelling that, given human beings as they are, no one can reasonably be asked to overcome his rancorous feelings' (p. 534). In the following section 81, on 'envy and equality', Rawls argues carefully that his 'principles of justice are not likely to arouse excusable... envy... to a troublesome extent' (p. 537). Thereafter, he discusses the conservative contention 'that the tendency to equality in modern social movements is the expression of envy' (p. 538), and Freud's lamentable suggestion that an egalitarian sense of justice is but an adult manifestation of childish feelings of envy and jealousy. Recently, indeed, a particular progressive tax in West Germany has been labelled an 'envy tax' *(Neidsteuer)*, as noted by Bös and Tillmann (1985). Anyway, Rawls is careful to distinguish 'rancorous' envy from the justifiable feelings of resentment at being treated unjustly. While envy may often form the basis of an appeal to justice, the claims that all appeals to justice rely on envy often fail to distinguish envy from resentment. I shall return to this later.

Equity as Absence of 'Envy'

More recently, however, 'envy' has acquired a precise technical sense in economic theory, following the (apparently independent) lead taken by Feldman and Kirman (1974) and by Varian (1974) in analysing a concept of 'equity' due to Foley (1967, p. 75). Apparently Foley was the first to use the term 'envy' in this sense, though only informally: Feldman, Kirman and Varian include it in their titles.

Consider any allocation (x^i_g) (g $= 1$ to n; $i = 1$ to m) of n goods to each of m individuals. Suppose these individuals have preferences represented by ordinal utility functions $U^i(x^i)$ ($i = 1$ to m) of each individual i's own (net) consumption vector x^i. Then individual i is said to *envy* j if $U^i(x^j) > U^i(x^i)$, so that i prefers j's allocation to his own. Notice that this *is* a purely technical definition; it tells us nothing about i's emotional or psychological state, whether i is unhappy because he prefers what j has, or whether i's 'envy' makes him want to harm j. There is no sin in this unemotional economists' concept of envy, but no particular ethical appeal either. Indeed, it might be better to say that 'i finds j's position to be *enviable*', to minimize the suggestions of emotion.

Nevertheless, Foley was concerned to introduce a concept of equity of welfare which overcomes the deficiencies of equality of after-tax income – deficiencies which are obvious when there are different public goods in different areas, different preferences for leisure as against consumption, and different needs as well. Thus Foley proposes the absence of 'envy' as a test of whether an allocation is equitable. Formally, (x^i_g) is *equitable* if $U^i(x^i) \geq U^i(x^j)$ for all pairs of individuals i and j.

Foley (1967) was careful to qualify this test. First, lifetime consumption plans must be considered so that the prodigal do not envy the higher later consumption enjoyed by the thrifty. Second, as he says:

> if a gas station attendant has the desire to be a painter but not the ability, it may be necessary to make the painter's life very unattractive in other ways before the gas station attendant will prefer his own; so unattractive, perhaps, that the painter

> will envy the attendant while the attendant is still envying him. These cases must be interpreted flexibly; either equivalents to the talents must be postulated which the gas station attendant does possess, or reasonable alternatives framed that abstract from the glamour and prestige of certain activities (p. 75).

Foley (1967, p. 76) concludes his discussion of 'equity' as follows:

> If tastes differ greatly, there is very little gained by the analysis since a very wide range of allocations will meet the equity criterion. The definition is offered only as a tentative contribution to a difficult and murky subject and concludes the sketchy discussion this paper will make of welfare economics.

Fairness and Other Extensions of Equity

In a one-good problem of dividing a cake, procedures for achieving 'fair' allocations, without envy and with no cake wasted, were discussed in works such as Steinhaus (1948, 1949), and Dubins and Spanier (1961) before Foley. Fairness with many goods was considered by Schmeidler and Vind (1972), by adding Pareto efficiency to the requirement that nobody should envy anybody else's *net trade vector* (as opposed to the consumption vector, which includes the endowment). Pazner and Schmeidler (1974) and Varian (1974) then came up with examples of economies with production in which there are no fair allocations, because Pareto efficiency requires skilled workers to supply more hours of labour than unskilled workers, and tastes are such that no allocation of consumption then avoids envy. Feldman and Kirman (1974) considered reducing the degree of envy, whereas Pazner and Schmeidler (1978) weakened the notion of equity to 'egalitarian equivalence' – finding an allocation (x^i_g) in which each individual i is indifferent between x^i and a consumption in an 'egalitarian' allocation $(\bar{x}^i{}_g)$ with $\bar{x}^i$ ndependent of i. Of course, in this egalitarian allocation there is no envy. These later developments all seem like attempts to rescue a dubious notion of equality without giving up first-best Pareto efficiency, even though that is surely unattainable anyway in economies with private information.

Envy and Resentment Reconsidered

In the definition of envy, each individual i compares the consumption vector x^j of another with his own, x^i, using i's own utility function U^i. But $U^i(x^j) > U^i(x^i)$ is insufficient for i's envy to be excusable, in Rawls's sense. Indeed, if x^j is preferable for i because j has some special needs met, i's envy is quite unjustifiable. As Sen (1970, ch. 9) points out in his discussion of Suppes's (1966) grading principles of justice, comparisons of x^j and x^i must allow for differences in tastes, needs, and so on. Thus the appropriate comparison in determining what is inequitable is rather whether $U^j(x^j) > U^i(x^i)$. If this is true, we might say that i *resents* j. Absence of resentment then requires all individuals to have equal utility levels; of course, this requires interpersonal comparisons of utility levels, of the kind used to make decisions 'in an original position', behind the 'veil of ignorance', before each individual knows his tastes. The technical sense of envy defined earlier differs from this technical notion of resentment precisely because it ignores the original position; not surprisingly, then, envy has no moral force, whereas resentment may well have.

Complete absence of resentment in this sense is probably too strong; but one should look for there to be no resentment in the weaker sense that no individual can legitimately feel treated unjustly by the institutions that determine his welfare. That, of course, reverts to Rawls (1971), though not necessarily to his particular concept of justice.

See Also

- ▶ Altruism
- ▶ Equity
- ▶ Fairness

Bibliography

Bös, D., and G. Tillmann. 1985. An 'Envy Tax': Theoretical principles and applications to the German surcharge on the rich. *Public Finance/Finances Publiques* 40: 35–63.

Brennan, G. 1973. Pareto desirable redistribution: The case of malice and envy. *Journal of Public Economics* 2: 173–183.
Chaudhuri, A. 1985. Formal properties of interpersonal envy. *Theory and Decision* 18: 301–312.
Dubins, L.E., and E.H. Spanier. 1961. How to cut a cake fairly. *American Mathematical Monthly* 68: 1–17.
Feldman, A., and A. Kirman. 1974. Fairness and envy. *American Economic Review* 64: 995–1005.
Foley, D.K. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7: 45–198.
Nozick, R. 1974. *Anarchy, state, and utopia*. New York: Basic Books.
Pazner, E., and D. Schmeidler. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41: 441–443.
Pazner, E., and D. Schmeidler. 1978. Egalitarian equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92: 671–687.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA/Oxford: Harvard University Press/Clarendon Press.
Schmeidler, D., and K. Vind. 1972. Fair net trades. *Econometrica* 40: 637–642.
Sen, A.K. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.
Steinhaus, H. 1948. The problem of fair division. *Econometrica* 16: 101–104.
Steinhaus, H. 1949. Sur la division pragmatique. *Econometrica* 17: 315–319.
Suppes, P. 1966. Some formal models of grading principles. *Synthese* 16: 284–306.
Varian, H.R. 1974. Equity, envy and efficiency. *Journal of Economic Theory* 9: 63–91.
Veblen, T. 1899. *The Theory of the leisure class*, 1967. New York: Macmillan and Viking.

Ephémérides du citoyen ou chronique de l'esprit National

Peter Groenewegen

Keywords

Baudeau, N.; Du Pont de Nemours, P.S.; *Ephémérides du Citoyen*; Mirabeau, V.R. Marquis de; Physiocracy

JEL Classifications

A1

French economic periodical issued in three series under different names from 1766 to 1772, 1774 to 1776 and in 1788. Published first as a bimonthly by its founder and first editor, l'Abbé Baudeau, it became a monthly as from January 1767 after Baudeau's conversion to Physiocracy by Mirabeau and Le Trosne. Its contents included contributed articles on economic and political subjects, book reviews, comments and letters to the editor, together with a chronicle of public events of interest to its readership. This provided its format from January 1769, when Du Pont de Nemours took over the editorship. Although censorship problems troubled the journal persistently (as disclosed in the Turgot–Du Pont correspondence, for this reason many issues appeared well after the ostensible month of publication) the first series was terminated by l'Abbé Terray in November 1772, presumably because it contained much vigorous criticism of his abolition of domestic free trade in grain. The first series produced therefore six issues in 1766 as a bi-monthly and 63 monthly issues from January 1767 to March 1772 inclusive. Under the title *Nouvelles Ephémérides ou Bibliothèque raisonnée de l'histoire, de la morale et de la politique*, it was revived by Baudeau after Turgot became Contrôleur-général in 1774, publishing 18 issues in all from January 1775 to June 1776, that is, the month after Turgot's dismissal from the ministry. A third series, *Nouvelles Ephémérides économiques* published three issues from January to March 1788, again under Baudeau's editorship, but his failing mental powers were presumably the reason why this final series ended so quickly.

Although initially set up by Baudeau in imitation of the English *Spectator*, within a year of its inception economics began to dominate its contents and many of the leading Physiocrats, in particular Mirabeau, Baudeau and Du Pont de Nemours, contributed most of the articles. A detailed discussion of its contents is given in Bauer (1894) and in Coquelin and Guillaumin (1854, pp. 710–12). Perhaps the most important piece it contained is Turgot's *Réflexions sur la formation et distribution des richesses* in serial form (*Ephémérides,* 1769, No. 11, pp. 12–56; No. 12, pp. 31–98; and 1770, No. I,

pp. 113–73), although with considerable unauthorized alterations and notes by Du Pont (see Groenewegen 1977, pp. xix–xxi). It also published foreign contributions in French translation, including Beccaria's inaugural lecture with copious notes and comments by Du Pont (*Ephémérides,* 1769, No. 6, pp. 57–152) and a contribution by Franklin on the increasing troubles between England and her American colonies (*Ephémérides,* 1768, No. 8, pp. 159–92). As an early, if not the first, economic journal, the *Ephémérides* remains an important part of economic literature and an indispensable source for those interested in the study of Physiocracy.

See Also

► Physiocracy

Bibliography

Bauer, S. 1894. Ephémérides. In *Dictionary of political economy*, vol. 1, ed. R.H.I. Palgrave. London: Macmillan.

Coquelin, C. and Guillaumin, H., ed. 1854. Ephémérides. In *Dictionnaire de l'économie politique*, vol. 1. Paris: Guillaumin, Hachette.

Groenewegen, P.D. 1977. *The economics of A.RJ. Turgot*. The Hague: Martinus Nijhoff.

Epistemic Game Theory: An Overview

Adam Brandenburger

Keywords

Common knowledge; Complete information; Epistemic game theory; Epistemic game theory: an overview; Game theory; Harsanyi, J. C; Incomplete information; Morgenstern, O; Non-cooperative game theory; Rational behaviour; Uncertainty; von Neumann, J

JEL Classification

C7

The following three articles survey some aspects of the foundations of noncooperative game theory. The goal of work in foundations is to examine in detail the basic ingredients of game analysis.

The starting point for most of game theory is a 'solution concept' – such as Nash equilibrium or one of its many variants, backward induction, or iterated dominance of various kinds. These are usually thought of as the embodiment of 'rational behaviour' in some way and used to analyse game situations.

One could say that the starting point for most game theory is more of an endpoint of work in foundations. Here, the primitives are more basic. The very idea of rational – or irrational – behaviour needs to be formalized. So does what each player might know or believe about the game – including about the rationality or irrationality of other players. Foundational work shows that even what each player knows or believes about what other players know or believe, and so on, can matter.

Investigating the basis of existing solution concepts is one part of work in foundations. Other work in foundations has uncovered new solution concepts with useful properties. Still other work considers changes even to the basic model of decision making by players – such as departures from the expected utility model or reasoning in various formal logics.

The first article, epistemic game theory: beliefs and types, by Marciano Siniscalchi, describes the formalism used in most work on foundations. This is the 'types' formalism going back to Harsanyi (1967–8). Originally proposed to describe the players' beliefs about the structure of the game (such as the payoff functions), the types approach is equally suited to describing beliefs about the play of the game or beliefs about both what the game is and how it will be played. Indeed, in its most general form, the formalism is simply a way to describe any multi-person uncertainty. Harsanyi's conception of a 'type' was a crucial breakthrough in game theory. Still, his work left many fundamental questions about multi-person

uncertainty unanswered. Siniscalchi's article surveys these later developments.

The second and third articles apply these tools to the two kinds of uncertainty mentioned. The second article, epistemic game theory: complete information, concerns the case where the matrix or tree itself is 'transparent' to the players, and what is uncertain are the actual strategies chosen by the players. The third article, epistemic game theory: incomplete information, by Aviad Heifetz, has the opposite focus: it covers the case of uncertainty about the game itself. (Following Harsanyi, the third article focuses on uncertainty about the payoffs, in particular.)

Both cases are important to the foundations programme. Because Nash equilibrium is 'as if' each player is certain (and correct) about the strategies chosen by the other players (Aumann and Brandenburger 1995, Section 7h), uncertainty of the first kind has played a small role in game theory to date. Uncertainty of the second kind is the topic of the large literatures on information asymmetries, incentives, and so on.

Interestingly, though, von Neumann and Morgenstern (1944) already appreciated the significance of both complete and incomplete information environments. Indeed, they asserted that phenomena often thought to be characteristic of incomplete-information settings could, in fact, arise in complete-information settings (1944, p. 31):

> Actually, we think that our investigations – although they assume 'complete information' without any further discussion – do make a contribution to the study of this subject. It will be seen that many economic and social phenomena which are usually ascribed to the individual's state of 'incomplete information' make their appearance in our theory and can be satisfactorily interpreted with its help.

This is indeed true, as work in the modern foundations programme shows. (Some instances are mentioned in what follows.) Overall, the foundations programme aims at a 'neutral' and comprehensive treatment of all ingredients of a game.

See Also

- ▶ Epistemic Game Theory: Beliefs and Types
- ▶ Epistemic Game Theory: Complete Information
- ▶ Epistemic Game Theory: Incomplete Information
- ▶ Game Theory
- ▶ Nash Equilibrium, Refinements of

My thanks to Rena Henderson and Michael James. The Stern School of Business provided financial support.

Bibliography

Aumann, R., and A. Brandenburger. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63: 1161–1180.

Harsanyi, J. 1967–8. Games with incomplete information played by 'Bayesian' players, I–III. *Management Science* 14, 159–182, 320–334, 486–502.

Von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press. 60th anniversary edn, 2004.

Epistemic Game Theory: Beliefs and Types

Marciano Siniscalchi

Abstract

Modelling what each agent believes about her opponents, what she believes her opponents believe about her, and so on, plays a prominent role in game theory and its applications. This article describes Harsanyi's formalism of type spaces, which provides a simple, elegant representation of probabilistic belief hierarchies. A special emphasis is placed on the construction of rich type spaces, which can generate all 'reasonable' belief hierarchies in a given game. Recent developments, employing richer representation of beliefs, are also considered.

Keywords

Belief hierarchies; Common knowledge; Epistemic game theory: beliefs and types; Harsanyi, J.C.; Kolmogorov's extension theorem; Monotonicity; Polish spaces; Preferences; Recursive preferences; Type spaces; Universal type space

JEL Classifications
C7

John Harsanyi (1967–8) introduced the formalism of type spaces to provide a simple and parsimonious representation of belief hierarchies. He explicitly noted that his formalism was not limited to modelling a player's beliefs about payoff-relevant variables: rather, its strength was precisely the ease with which Ann's beliefs about Bob's beliefs about payoff variables, Ann's beliefs about Bob's beliefs about Ann's beliefs about payoff variables, and so on, could be represented.

This feature plays a prominent role in the epistemic analysis of solution concepts (see epistemic game theory: complete information), as well as in the literature on global games (Morris and Shin 2003) and on robust mechanism design (Bergemann and Morris 2005). All these applications place particular emphasis on the expressiveness of the type-space formalism. Thus, a natural question arises: just how expressive is Harsanyi's approach?

For instance, solution concepts such as Nash equilibrium or rationalizability can be characterized by means of restrictions on the players' mutual beliefs. In principle, these assumptions could be formulated directly as restrictions on players' hierarchies of beliefs; but in practice the analysis is mostly carried out in the context of a type space à la Harsanyi. This is without loss of generality only if Harsanyi type spaces do not themselves impose restrictions on the belief hierarchies that can be represented. Similar considerations apply in the context of robust mechanism design.

A rich literature addresses this issue from different angles, and for a variety of basic representations of beliefs. This article focuses on hierarchies of probabilistic beliefs; however, some extensions are also mentioned. For simplicity, attention is restricted to two players, denoted '1' and '2' or 'i' and '$-i$.'

Probabilistic Type Spaces and Belief Hierarchies

Begin with some mathematical preliminaries. A topology on a space X is deemed Polish if it is separable and completely metrizable; in this case, X is itself deemed a Polish space. Examples include finite sets, Euclidean space $\mathbb{R}^n$ and closed subsets thereof. A countable product of Polish spaces, endowed with the product topology, is itself Polish. For any topological space X, the notation $\Delta(X)$ indicates the set of Borel probability measures on X. If the topology on X is Polish, then the weak* topology on $\Delta(X)$ is also Polish (for example, Aliprantis and Border 1999, Theorem 14.15). A sequence $\{\mu^k\}k \geq 1$ in $\Delta(X)$ converges in the weak* sense to a measure $\mu \in \Delta(x)$, written μ^k *LongRightArrow*;$^{w^*}$ μ, if and only if, for every bounded, continuous function $\psi : X \to \mathbb{R}$, $\int_X \psi d\mu^k \to \int_X \psi d\mu$. The weak* topology on $\Delta(X)$ is especially meaningful and convenient when X is a Polish space: see Aliprantis and Border (1999, Chapter 14) for an overview of its properties. Finally, if μ is a measure on some product space $X \times Y$, the marginal of μ on X is denoted $\text{marg}_x\mu$.

The basic ingredient of the players' hierarchical beliefs is a description of payoff-relevant or fundamental uncertainty. Fix two sets S_1 and S_2, hereinafter called the *uncertainty domains*; the intended interpretation is that S_{-i} describes aspects of the strategic situation that Player i is uncertain about. For example, in an independent private-values auction, each set S_i could represent bidder i's possible valuations of the object being sold, which is not known to bidder $-i$. In the context of interactive epistemology, S_i is usually taken to be Player i's strategy space. It is sometimes convenient to let $S_1 = S_2 \equiv S$; in this case, the formalism introduced below enables one to formalize the assumption that each player observes different aspects of the common uncertainty domain S (for instance, different signals correlated with the common, unknown value of an object offered for sale).

An (S_1, S_2)-*based type space* is a tuple $J = (T_i, g_i)_{i=1,2}$ such that, for each $i = 1, 2$, T_i is a Polish space and $g_i : T_i \to \Delta(S_{-I} \times T_{-i})$ is continuous. As noted above, type spaces can represent hierarchies of beliefs; it is useful to begin with an example. Let $S_1 = S_2 = \{a, b\}$ and consider the type space defined in Table 1. To interpret, for every $i = 1,2$, the entry in the row

Epistemic Game Theory: Beliefs and Types, Table 1 A type space

T_1	a, t_2	a, t'_2	b, t_2	b, t'_2
t_1	1	0	0	0
t'_1	0	0.3	0	0.7
T_2	a, t_1	a, t'_1	b, t_1	b, t'_1
t_2	0	0.5	0.5	0
t'_2	0	0	0	1

corresponding to t_i and (s_{-i}, t_{-i}) is $g_i(t_i)(\{(s_{-i}, t_{-i})\})$. Thus, for instance, $g_1(t_1)\left(\left\{\left(a, t'_2\right)\right\}\right) = 0$; $g_2(t_2)(\{b\} \times T_1) = 0.5$.

Consider type t_1 of Player 1. She is certain that $s_2 = a$; furthermore, she is certain that Player 2 believes that $s_1 = a$ and $s_1 = b$ are equally likely. Taking this one step further, type t_1 is certain that Player 2 assigns probability 0.5 to the event that Player 1 believes that $s_2 = b$ with probability 0.7.

These intuitive calculations can be formalized as follows. Fix an (S_1, S_2)-based type space $J = (T_i, g_i)_{i=1,2}$; for every $i = 1, 2$ define the set X^0_{-i} and the function $h^1_i : T_i \to \Delta\left(X^{-i}_0\right)$ by

$$X^0_{-i} = S_{-i} \text{ and } \forall t_i \in T_i, \; h^1_i(t_i) = \text{marg}_{S_{-i}} g_i(t_i). \tag{1}$$

Thus, $h^1_i(t_i)$ represents the *first-order beliefs* of type t_i in type space T – her beliefs about the uncertainty domain S_{-i}. Note that each $X^0_{-i} = S_{-i}$ is Polish. Proceeding inductively, assuming that $X^0_{-i}, \ldots, X^{k-1}_{-i}$ and $h^1_i, \ldots, h^k_i$ have been defined up to some $k > 0$ for $i = 1,2$, and that all sets X^l_{-i}, $l = 0, \ldots, k-1$ are Polish, define the set X^k_{-i} and the functions $h^{k+1}_i : T_i \to \Delta\left(X^k_{-i}\right)$ for $i = 1, 2$ by

$$X^k_{-i} = X^{k-1}_{-i} \times \Delta\left(X^{k-1}_i\right) \text{ and } \forall t_i \in T_i, \; h^{K+1}_i(t_i)(E) = g_i(t_i)\left(\left\{(s_{-i}, t_{-i}) \in S_{-i} \times T_{-i} : \left(s_{-i}, h^k_{-i}(t_{-i})\right) \in E\right\}\right) \tag{2}$$

for every Borel subset E of X^k_{-i}. Thus, $h^2_1(t_1)$ represents the *second-order beliefs* of type t_1 – her beliefs about *both* the uncertainty domain $S_2 = X^0_2$ and Player 2's beliefs about S_1, which by definition belong to the set $\Delta\left(X^0_1\right) = \Delta(S_1)$. Similarly, $h^{k+1}_i(t_i)$ represents type t_i's $(k+1)$-th order beliefs.

Observe that type t_i's second-order beliefs are defined over $X^0_2 \times \Delta\left(X^0_1\right) = S_2 \times \Delta(S_1)$, rather than just over $\Delta\left(X^0_1\right) = \Delta(S_1)$; a similar statement holds for her $(k+1)$-th order beliefs. This is crucial in many applications. For instance, a typical assumption in the literature on epistemic foundations of solution concepts is that Player 1 believes that Player 2 is rational. Letting S_i be the set of actions or strategies of Player i in the game under consideration, this can be modelled by assuming that the support of $h^2_1(t_1)$ consists of pairs $(s_2, \mu_1) \in S_2 \times \Delta(S_1)$ wherein s_2 is a best response to μ_1. Clearly, such an assumption could not be formalized if $h^2_1(t_1)$ only conveyed information about type t_1's beliefs on Player 2's first-order beliefs: even though type t_1's beliefs about the action played by Player 2 could be retrieved from $h^1_1(t_1)$, it would be impossible to tell whether each action that type t_1 expects to be played is matched with a belief that rationalizes it.

Note that, since X^{k-1}_i and X^{k-1}_{-i} are assumed Polish, so are $\Delta\left(X^{k-1}_i\right)$ and X^k_{-i}. Also, each function h^k_i is continuous.

Finally, it is convenient to define a function that associates to each type $t_i \in T_i$ an entire *belief hierarchy*: to do so, define the set H_i and, for $i = 1, 2$, the function $h_i : T_i \to H_i$ by

$$H_i = \prod_{k \geq 0} \Delta\left(X^k_{-i}\right) \text{ and } \forall t_i \in T_i, \; h_i(t_i) = \left(h^1_i(t_i), \ldots, h^{k+1}_i(t_i), \ldots\right). \tag{3}$$

Thus, H_i is the set of all hierarchies of beliefs; notice that, since each X^k_{-i} is Polish, so is H_i.

Rich Type Spaces

The preceding construction suggests a rather direct way to ask how expressive Harsanyi's

notion of a type space is: can one construct a type space that generates *all* hierarchies in H_i?

A moment's reflection shows that this question must be refined. Fix a type space $(T_i, g_i)_{i=1,2}$ and a type $t_i \in T_i$; recall that, for reasons described above, the first- and second-order beliefs of type t_i satisfy $h_i^1(t_i) \in \Delta(S_{-i})$ and $h_i^2(t_i) \in \Delta\left(X_{-i}^0 \times \Delta\left(X_i^0\right)\right) = \Delta(S_{-i} \times \Delta(S_i))$ respectively. This, however, creates the potential for redundancy or even contradiction, because both $h_i^1(t_i)$ and $\text{marg}_{S_{-i}} h_i^2(t_i)$ can be viewed as 'type t_i's beliefs about S_{-i}'. A similar observation applies to higher-order beliefs. Fortunately, it is easy to verify that, for every type space $(T_i, g_i)_i = 1,2$ and type $t_i \in T_i$, the following *coherency* condition holds:

$$\forall k > 1, \ \text{marg}_{X_{-i}^{k-2}} h_i^k(t_i) = h_i^{k-1}(t_i); \quad (4)$$

To interpret, recall that $h_i^k(t_i) \in \left(X_{-i}^{k-1}\right) = \Delta\left(X_{-i}^{k-2} \times \Delta\left(X_{-i}^{k-2}\right)\right)$. Thus, in particular, $\text{marg}_{S_{-i}} h_i^2(t_i) = h_i^1(t_i)$.

Since H_i is defined as the set of *all* hierarchies of beliefs for Player i, some (in fact, 'most') of its elements are not coherent. As noted above, no type space can generate incoherent hierarchies; more importantly, coherency can be viewed as an integral part of the interpretation of interactive beliefs. How could an individual simultaneously hold (infinitely) many distinct first-order beliefs? Which of these should be used, say, to verify whether she is rational? This motivates restricting attention to coherent hierarchies, defined as follows:

$$H_i^c = \left\{ \left(\mu_i^1, \ \mu_i^2, \ldots\right) \in H_i : \forall k > 1, \ \text{marg}_{X_{-i}^{k-2}} \mu_i^k = \mu_i^{k-1} \right\}. \quad (5)$$

Since $\text{marg}_{X_{-i}^{k-2}} : \Delta\left(X_{-i}^{k-1}\right) \to \Delta\left(X_{-i}^{k-2}\right)$ is continuous, H_i^c is a closed, hence Polish subspace of H_i.

Brandenburger and Dekel (1993, Proposition 1) show that there exist homeomorphisms $g_i^c : H_i^c \to \Delta(S_{-i} \times H_{-i})$: that is, every coherent hierarchy corresponds to a distinct belief over the uncertainty domain and the hierarchies of the opponent, and conversely. Furthermore, this homeomorphism is canonical, in the following sense. Note that $S_{-i} \times H_{-i} = S_{-i} \times \Pi_{k \geq 0} \Delta\left(K_i^k\right) = K_{-i}^k \times \Pi_{l>k} \Delta\left(X_i^l\right)$.

Then it can be shown that, if $\mu_i = \left(\mu_i^1, \ \mu_i^2, \ldots\right) \in H_i^c$, then $\text{marg}_{X_{-i}^k} g_i^c(\mu_i) = \mu_i^{k+1}$ Intuitively, the marginal belief associated with μi over the first k orders of the opponent's beliefs is precisely what it should be, namely μ_i^{k+1}. The proof of these results builds upon Kolmogorov's extension theorem, as may be suggested by the similarity of the coherency condition in Eq. (5) with the notion of Kolmogorov consistency: cf. for example Aliprantis and Border (1999, Theorem 14.26).

This result does not quite imply that all coherent hierarchies can be generated in a suitable type space; however, it suggests a way to obtain this result. Notice that the belief on $S_{-i} \times H_{-i}$ associated by the homeomorphism g_i^c to a coherent hierarchy μ_i may include *in*coherent hierarchies $\nu_{-i} \in H_{-i} / H_{-i}^c$ in its support. This can be interpreted in the following terms: if Player i's hierarchical beliefs are given by μ_i, then she is coherent, but she is not certain that her opponent is. On the other hand, consider a type space $(T_i, g_i)_{i=1,2}$; as noted above, for every player i, each type $t_i \in T_i$ generates a coherent hierarchy $h_i(t_i) \in H_i^c$. So, for instance, if (s_1, t_1) is in the support of $g_2(t_2)$ then t_1 also generates a coherent hierarchy. Thus, not only is type t_2 of Player 2 coherent: he is also certain (believes with probability one) that Player 1 is coherent. Iterating this argument suggests that *hierarchies of beliefs generated by type spaces display common certainty of coherency.*

Motivated by these considerations, let

$$H_i^0 = H_i^c \text{ and } \forall k > 0,$$
$$H_i^k = \left\{ \mu_i \in H_i^{k-1} : g_i^c(\mu_i)\left(S_{-i} \times H_{-i}^{k-1}\right) = 1 \right\}. \quad (6)$$

Thus, H_i^0 is the set of coherent hierarchies for Player i; H_i^1 is the set of hierarchies that are coherent and correspond to beliefs that display certainty of the opponent's coherency; and so on. Finally, let $H_i^* = \cap_{k \geq 0} H_i^k$. Each element of H^* is intuitively consistent with coherency and common certainty of coherency.

E

Brandenburger and Dekel (1993, Proposition 2) show that the restriction g_i^* of g_i^c to H_i^* is a homeomorphism between H^*_i and $\Delta\left(S_{-i} \times H^*_{-i}\right)$; furthermore, it is canonical in the sense described above. This implies that the tuple $\left(H_i^*, g_i^*\right)_{i=1,2}$ is a type space in its own right – the (S_1, S_2)-based *universal type space*.

The existence of a universal type space fully addresses the issue of richness. Since the homeomorphism g_i^* is canonical, it is easy to see that the hierarchy generated as per Eqs. (1) and (2) by any 'type' $t_i = (\mu^1, \mu^2, s) \in H_i^*$ in the universal type space $\left(H_i^*, g_i^*\right)_{i=1,2}$ is t_i itself; thus, since H_i^* *consists* of all hierarchies that are coherent and display common certainty of consistency, the universal type space also *generates* all such hierarchies.

The type space $\left(H_i^*, g_i^*\right)_{i=1,2}$ is rich in two additional, related senses. First, as may be expected, every belief hierarchy for Player i generated by an arbitrary type space is an element of H^*; this implies that every type space $(T_i, g_i)_{i=1,2}$ can be uniquely embedded in $\left(H_i^*, g_i^*\right)_{i=1,2}$ as a 'belief-closed' subset: see Battigalli and Siniscalchi (1999, Proposition 8.8). Call a type space *terminal* if, like $\left(H_i^*, g_i^*\right)_{i=1,2}$, it embeds all other type spaces as belief-closed subsets.

Second, since each function g_i^* is a homeomorphism, in particular it is a surjection (that is, onto). Call a type space $(Ti, gi)_{i=1,2}$ *complete* if every map g_i is onto. (This should not be confused with the topological notion of completeness.) Thus, the universal type space $\left(H_i^*, g_i^*\right)_{i=1,2}$ is complete. It is often the case that, when a universal type space is employed in the epistemic analysis of solution concepts, the objective is precisely to exploit its completeness. Furthermore, for certain representations of beliefs, it is not known whether universal type spaces can be constructed; however, the existence of complete type spaces can be established, and is sufficient for the purposes of epistemic analysis. The next section provides examples.

Alternative Constructions and Extensions

The preceding discussion adopts the approach proposed by Brandenburger and Dekel (1993), which has the virtue of relying on familiar ideas from the theory of stochastic processes. However, the first constructions of universal type spaces consisting of hierarchies of beliefs are due to Armbruster and Böge (1979), Böge and Eisele (1979) and Mertens and Zamir (1985).

From a technical point of view, Mertens and Zamir (1985) assume that the state space S is compact Hausdorff and beliefs are regular probability measures. Heifetz and Samet (1998b) instead drop topological assumptions altogether: in their approach, both the underlying set of states and the sets of types of each player are modelled as measurable spaces. They show that a terminal type space can be explicitly constructed in this environment.

In all the contributions mentioned so far, beliefs are modelled as countably additive probabilities. The literature has also examined other representations of beliefs, broadly defined.

A *partitional structure* (Aumann 1976) is a tuple $(\Omega,(\sigma_i, P_i)_{i = 1,2})$, where Ω is a (typically finite) space of 'possible worlds', every $\sigma_i : \Omega \to S_i$ indicates the realization of the basic uncertainty corresponding to each element of Ω, and every P_i is a partition of Ω. The interpretation is that, at any world $\omega \in \Omega$, Player i is only informed that the true world lies in the cell of the partition P_i containing ω, denoted $P_i(\omega)$. The *knowledge operator* for Player i can then be defined as

$$\forall E \subset \Omega, K_i(E) = \{\omega \in \Omega : P_i(\omega) \subseteq E\}.$$

Notice that no probabilistic information is provided in this environment (although it can be easily added).

Heifetz and Samet (1998a) show that a terminal partitional structure does not exist. This result was extended to more general 'possibility' structures by Meier (2005). Brandenburger and Keisler (2006) establish related non-existence results for complete structures. However, recent contributions show that topological assumptions, which play a key role in the constructions of Mertens and Zamir (1985) and Brandenburger and Dekel (1993), can also deliver existence results in non-probabilistic settings. For instance, Mariotti

et al. (2005) construct a structure that is universal, complete and terminal for possibility structures.

Other authors investigate richer probabilistic representations of beliefs. Battigalli and Siniscalchi (1999) construct a universal, terminal, and complete type space for *conditional probability system*, or collections of probability measures indexed by relevant conditioning events (such as histories in an extensive game) and related by a version of Bayes's rule. This type space is used in (2002) to provide an epistemic analysis of forward induction. Brandenburger et al. (2006) construct a complete type space for *lexicographic sequences*, which may be thought of as an extension of lexicographic probability systems (Blume et al. 1991) for infinite domains. They then use it to provide an epistemic characterization of iterated admissibility.

Non-probabilistic representations of beliefs that reflect a concern for ambiguity (Ellsberg, 1961) have also been considered. Heifetz and Samet (1998b) observe that their measure-theoretic construction extends to beliefs represented by continuous *capacities*, that is non-additive set functions that preserve monotonicity with respect to set inclusion. Motivated by the multiple-priors model of Gilboa and Schmeidler (1989), Ahn (2007) constructs a universal type space for sets of probabilities.

Epstein and Wang (1996) approach the richness issue taking *preferences*, rather than beliefs, as primitive objects. In their setting, an S-based type space is a tuple $(T_i, g_i)_{i=1,2}$, where, for every type t_i, $g_i(t_i)$ is a suitably regular preference over *acts* defined on the set $S \times T_{-i}$. The analysis in the preceding section can be viewed as a special case of Epstein and Wang (1996), where preferences conform to expected-utility theory. Epstein and Wang construct a universal type space in this framework (see also Di Tillio 2006).

Finally, constructions analogous to that of a universal type space appear in other, unrelated contexts. For instance, Epstein and Zin (1989) develop a class of recursive preferences over infinite-horizon temporal lotteries; to construct the domain of such preferences, they employ arguments related to Mertens and Zamir's. Gul and Pesendorfer (2004) employ analogous techniques to analyse self-control preferences over infinite-horizon consumption problems.

See Also

- ▶ Epistemic Game Theory: An Overview
- ▶ Epistemic Game Theory: Complete Information
- ▶ Epistemic Game Theory: Incomplete Information

E

Bibliography

Ahn, D. 2007. Hierarchies of ambiguous beliefs. *Journal of Economic Theory* 136: 286–301.

Aliprantis, C., and K. Border. 1999. *Infinite dimensional analysis*. 2nd ed. Berlin: Springer.

Armbruster, W., and W. Böge. 1979. Bayesian game theory. In *Game theory and related topics*, ed. O. Moeschlin and D. Pallaschke. Amsterdam: North-Holland.

Aumann, R. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1236–1239.

Battigalli, P., and M. Siniscalchi. 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory* 88: 188–230.

Battigalli, P., and M. Siniscalchi. 2002. Strong belief and forward induction reasoning. *Journal of Economic Theory* 106: 356–391.

Bergemann, D., and S. Morris. 2005. Robust mechanism design. *Econometrica* 73: 1521–1534.

Blume, L., A. Brandenburger, and E. Dekel. 1991. Lexicographic probabilities and choice under uncertainty. *Econometrica* 59: 61–79.

Böge, W., and T. Eisele. 1979. On solutions of Bayesian games. *International Journal of Game Theory* 8: 193–215.

Brandenburger, A., and E. Dekel. 1993. Hierarchies of beliefs and common knowledge. *Journal of Economic Theory* 59: 189–198.

Brandenburger, A., Friedenberg, A. and Keisler, H.J. 2006. *Admissibility in games*. Unpublished, Stern School of Business, New York University.

Brandenburger, A., and J. Keisler. 2006. An impossibility theorem on beliefs in games. *Studia Logica* 84: 211–240.

Di Tillio, A. 2006. *Subjective expected utility in games*. Working Paper No. 311, IGIER, Università Bocconi.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Epstein, L., and T. Wang. 1996. Beliefs about beliefs without probabilities. *Econometrica* 64: 1343–1373.

Epstein, L., and S. Zin. 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* 57: 937–969.
Gilboa, I., and D. Schmeidler. 1989. Maxmin-expected utility with a non-unique prior. *Journal of Mathematical Economics* 18: 141–153.
Gul, F., and W. Pesendorfer. 2004. Self-control and the theory of consumption. *Econometrica* 72: 119–158.
Harsanyi, J. 1967–8. Games of incomplete information played by Bayesian players. Parts I, II, III. *Management Science* 14:159–182, 320–334, 486–502.
Heifetz, A., and D. Samet. 1998a. Knowledge spaces with arbitrarily high rank. *Games and Economic* 22: 260–273.
Heifetz, A., and D. Samet. 1998b. Topology-free typology of beliefs. *Journal of Economic Theory* 82: 324–381.
Mariotti, T., M. Meier, and M. Piccione. 2005. Hierarchies of beliefs for compact possibility models. *Journal of Mathematical Economics* 41: 303–324.
Meier, M. 2005. On the nonexistence of universal information structures. *Journal of Economic Theory* 122: 132–139.
Mertens, J.F., and S. Zamir. 1985. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14: 1–29.
Morris, S., and H. Shin. 2003. Global games: Theory and applications. In *Advances in economics and econometrics (Proceedings of the eight world congress of the econometric society)*, ed. M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge: Cambridge University Press.

Epistemic Game Theory: Complete Information

Adam Brandenburger

Abstract

The epistemic programme can be viewed as a methodical construction of game theory from its most basic elements – rationality and irrationality, belief and knowledge about such matters, beliefs about beliefs, knowledge about knowledge, and so on. To date, the epistemic field has been mainly focused on game matrices and trees – that is, on the non-cooperative branch of game theory. It has been used to provide foundations for existing non-cooperative solution concepts, and also to uncover new solution concepts. The broader goal of the programme is to provide a method of analysing different sets of assumptions about games in a precise and uniform manner.

Keywords

Admissibility; Backward induction; Common knowledge; Conditional probability systems; Correlation; Epistemic game theory; Epistemic game theory: complete information; Finite games; Invariance; Iterated dominance; Lexicographic probability systems; Rational behaviour; Rationalizability; Strong dominance; Type structures; Uncertainty; Weak dominance

JEL Classifications

C7

Epistemic Analysis

Under the epistemic approach, the traditional description of a game is augmented by a mathematical framework for talking about the rationality or irrationality of the players, their beliefs and knowledge, and related ideas.

The first step is to add sets of *types* for each of the players. The apparatus of types goes back to Harsanyi (1967–8), who introduced it as a way to talk formally about the players' beliefs about the payoffs in a game, their beliefs about other players' beliefs about the payoffs, and so on. (See epistemic game theory: incomplete information.) But the technique is equally useful for talking about uncertainty about the actual play of the game – that is, about the players' beliefs about the strategies chosen in the game, their beliefs about other players' beliefs about the strategies, and so on. This survey focuses on this second source of uncertainty. It is also possible to treat both kinds of uncertainty together, using the same technique.

We give a definition of a type structure as commonly used in the epistemic literature, and an example of its use.

Fix an n-player finite strategic-form game $\langle S^1, \ldots, S^n, \pi^1, \ldots, \pi^n \rangle$. Some notation: given sets $X^1,\ldots, X^n$, let $X = \times_{i=1}^n X^i$ and $X^{-i} = \times_{j \neq i} X^j$. Also, given a finite set Ω, write $\mathcal{M}(\Omega)$ for set of all probability measures on Ω.

Definition 1.1 *An $(S^1,\ldots, S^n)$-based **(finite) type structure** is a structure*

$$\langle S^1, \ldots, S^n; T^1, \ldots T^n; \lambda^1, \ldots \lambda^n \rangle,$$

*where each T^i is a finite set, and each $\lambda^i : T^i \to \mathcal{M}\left(S^{-i} \times T^{-i}\right)$. Members of T^i are called **types** for player i. Members of $S \times T$ are called **states (of the world)**.*

For some purposes – see, for example, sections "Conditions for Backward Induction and Conditions for Iterated Admissibility" – it is important to consider infinite type structures. Topological assumptions are then made on the type spaces T_i.

A particular state $(s^1, t^1,\ldots, s^n, t^n)$ describes the strategy chosen by each player, and also each player's type. Moreover, a type t^i for player i induces, via a natural induction, an entire hierarchy of beliefs – about the strategies chosen by the players $j \neq i$, about the beliefs of the players $j \neq i$, and so on. (See epistemic game theory: beliefs and types.)

The following example is similar to one in Aumann and Brandenburger (1995, pp. 1166–7).

Example 1.1 (A Coordination Game) *Consider the coordination game in* Fig. 1.1 *(where Ann chooses the row and Bob the column), and the associated type structure in* Fig. 1.2.

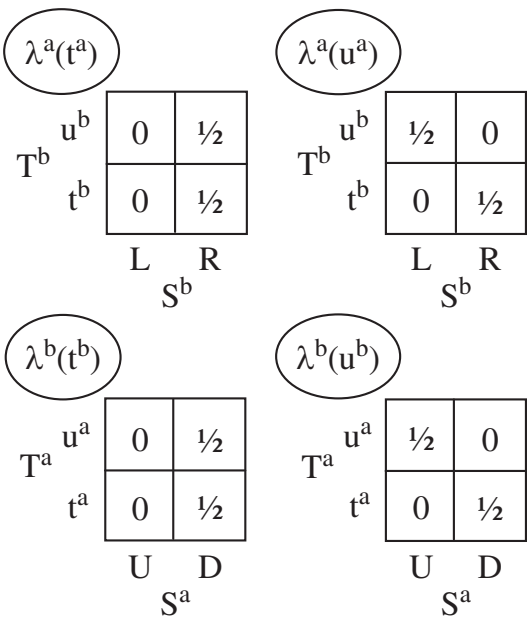
There are two types t^a, u^a for Ann, and two types t^b, u^b for Bob. The measure associated with each type is as shown. Fix the state (D, t^a, R, t^b). At this state, Ann plays D and Bob plays R. Ann is 'correct' about Bob's strategy. (Her type t^a assigns probability 1 to Bob's playing R.) Likewise, Bob is correct about Ann's strategy. Ann, though, thinks it possible Bob is wrong about her strategy. (Her type assigns probability 1/2 to type u^b for Bob, which assigns probability 1/2 to Ann's playing U, not D.) Again, likewise with Bob.

What about the rationality or irrationality of the players? At state (D, t^a, R, t^b), Ann is rational. Her strategy maximizes her expected payoff, given her first-order belief (which assigns probability 1 to R). Likewise, Bob is rational. Ann, though, thinks it possible Bob is irrational. (She assigns probability 1/2 to (R, u^b). With type u^b, Bob gets a higher expected payoff from L than R.) The situation with Bob is again symmetric.

Summing up, the example is just a description of a game situation, not a prediction. A type structure is a descriptive tool. Note, too, that the example includes both rationality and irrationality, and also allows for incorrect as well as correct beliefs (for example, Ann thinks it possible Bob is irrational, though in fact he isn't). These are typical features of the epistemic approach.

Epistemic Game Theory: Complete Information, Fig. 1.1

	L	R
U	2, 2	0, 0
D	0, 0	1, 1



Epistemic Game Theory: Complete Information, Fig. 1.2

Two comments on type structures. First, we can ask whether Definition 1.1 above is to be taken as primitive or derived. Arguably, hierarchies of beliefs are the primitive, and types are simply a convenient tool for the analyst. See epistemic game theory: beliefs and types for further discussion.

Second, note that Definition 1.1 applies to finite games. These will be the focus of this survey. There is nothing yet approaching a developed literature on epistemic analysis of infinite games.

Early Results

A major use of type structures is to identify conditions on the players' rationality, beliefs, and so on, that yield various solution concepts.

A very basic solution concept is iterated dominance. This involves deleting from the matrix all strongly dominated strategies, then deleting all strategies that become strongly dominated in the resulting submatrix, and so on until no further deletion is possible. (It is easy to check that in finite games – as considered in this survey – the residual set will always be non-empty.) Call the remaining strategies the *iteratively undominated* (*IU*) strategies. There is a basic equivalence: a strategy is not strongly dominated if and only if there is a probability measure on the product of the other players' strategy sets under which it is optimal. Using this, IU can also be defined as follows: delete from the matrix any strategy that isn't optimal under some measure on the product of the other players' strategy sets. Consider the resulting sub-matrix and delete strategies that don't pass this test on the sub-matrix, and so on.

The second definition suggests what a formal epistemic treatment of IU should look like. A rational player will choose a strategy which is optimal under some measure. This is the first round of deletion. A player who is rational and believes the other players are rational will choose a strategy which is optimal under a measure that assigns probability 1 to the strategies remaining after the first round of deletion. This gives the second round of deletion. And so on.

Type structures allow a formal treatment of this idea. First the formal definition of rationality. This is a property of strategy-type pairs. Say (s^i, t^i) is **rational** if s^i maximizes player i's expected payoff under the marginal on S^{-i} of the measure $\lambda^i(t^i)$.

Say type t^i of player i **believes** an event $E \subseteq S^{-i} \times T^{-i}$ if $\lambda^i(t^i)(E) = 1$, and write

$$B^i(E) = \left\{t^i \in T^i : t^i \text{ believes } E\right\}.$$

Now, for each player i, let R^i_1 be the set of all rational pairs (s^i, t^i), and for $m > 0$ define R^i_m inductively by

$$R^i_{m+1} = R^i_m \cap \left[S^i \times B^i\left(R^{-i}_m\right)\right].$$

Definition 2.1 *If* $(s^1, t^1, \ldots, s^n, t^n) \in R_{m+1}$, *say there is* ***rationality andm th-order elief of rationality (RmBR)*** *at this state. If* $(s^1, t^1, \ldots, s^n, t^n) \in \cap^{\infty}_{m=1} R_m$ *say there is* ***rationality and common belief of rationality (RCBR)*** *at this state.*

These definitions yield an epistemic characterization of IU: *Fix a type structure and a state* $(s^1, t^1, \ldots, s^n, t^n)$ *at which there is RCBR. Then the strategy profile* $(s^1, \ldots, s^n)$ *is IU. Conversely, fix an IU profile* $(s^1, \ldots, s^n)$. *There is a type structure and a state* $(s^1, t^1, \ldots, s^n, t^n)$ *at which there is RCBR.* Results like this can be found in the early literature – see, among others, Brandenburger and Dekel (1987) and Tan and Werlang (1988).

An important stimulus to the early literature was the pair of papers by Bernheim (1984) and Pearce (1984), which introduced the solution concept of *rationalizability*. This differs from IU by requiring on each round that a player's probability measure on the product of the other players' (remaining) strategy sets be a product measure – that is, be independent. Thus the set of rationalizable strategy profiles is contained in the IU set. It is well known that there are games (with three or more players) in which inclusion is strict.

The argument for the independence assumption is that in non-cooperative game theory it is supposed that players do not coordinate their strategy choices. Interestingly though, correlation is

consistent with the non-cooperative approach. This view is put forward in Aumann (1987). (Aumann, 1974, introduced the study of correlation into non-cooperative theory.) Consider an analogy to coin tossing. A correlated assessment over coin tosses is possible, if there is uncertainty over the coin's parameter or 'bias'. (The assessment is usually required to be conditionally i.i.d., given the parameter.) Likewise, in a game, Charlie might have a correlated assessment over Ann's and Bob's strategy choices, because, say, he thinks Ann and Bob have observed similar signals before the game (but is uncertain what the signal was).

The same epistemic tools used to understand IU can be used to characterize other solution concepts on the matrix. Aumann and Brandenburger (1995, Preliminary Observation) point out that pure-strategy Nash equilibrium is characterized by the simple condition that each player is rational and assigns probability 1 to the actual strategies chosen by the other players. (Thus, in Example 1.1 above, these conditions hold at the state (D, t^a, R, t^b), and (D, R) is indeed a Nash equilibrium.) As far as mixed strategies are concerned, in the epistemic approach to games these don't play the central role that they do under equilibrium analysis. Built into the set-up of section "Epistemic Analysis" is that each player makes a definite choice of (pure) strategy. (If a player does have the option of making a randomized choice, this can be added to the – pure – strategy set. Indeed, in a finite game, a finite number of such choices can be added.) It is the other players who are uncertain about this choice. Harsanyi (1973) originally proposed this shift in thinking about randomization. Aumann and Brandenburger (1995) give an epistemic treatment of mixed-strategy Nash equilibrium along these lines.

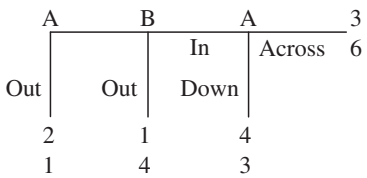
Aumann (1987) asks a question about an outside observer of a game. He provides conditions under which the observer's assessment of the strategies chosen will be the distribution of a correlated equilibrium (as defined in his 1974 paper). The distinctive condition in (1987) is the so-called Common Prior Assumption, which says that the probability assessment associated with each player's type is the same as the observer's assessment, except for being conditioned on what the type in question knows. A number of papers have investigated foundations for this assumption – see, among others, Morris (1994), Samet (1998), Bonanno and Nehring (1999), Feinberg (2000), Halpern (2002), and also the exchange between Gul (1998) and Aumann (1998).

Next Steps: The Tree

An important next step in the epistemic programme was extending the analysis to game trees. A big motivation for this was to understand the logical foundation of *backward induction* (*BI*). At first sight, BI is one of the easiest ideas in game theory. If Ann, the last player to move, is rational, she will make the BI choice. If Bob, the second-to-last player to move, is rational and thinks Ann is rational, he will make the choice that is maximal given that Ann makes the BI choice – that is, he too will make the BI choice. And so on back in the tree, until the BI path is a identified (Aumann, 1995).

For example, Fig. 3.1 is three-legged centipede (Rosenthal, 1981). (The top payoffs are Ann's, and the bottom payoffs are Bob's.) BI says Ann plays *Out* at her first node. But what if she doesn't? How will Bob react? Perhaps Bob will conclude that Ann is an irrational player, who plays *Across*. That is, Bob might play *In*, hoping to get a payoff of 6 (better than 4 from *Out*). Perhaps, anticipating this, Ann will in fact play *Down*, hoping to get 4 (better than 2 from playing *Out*).

Many papers have examined this conceptual puzzle with BI – see, among others, Binmore



Epistemic Game Theory: Complete Information, Fig. 3.1

(1987), Bicchieri (1988, 1989), Basu (1990), Bonanno (1991), and Reny (1992).

A key step in resolving the puzzle is extending the epistemic tools of section "Epistemic Analysis", to be able to talk formally about rationality, beliefs and so on in the tree.

Example 3.1 (Three-Legged Centipede) Figure 3.2 *is a type structure for three-legged Centipede.*

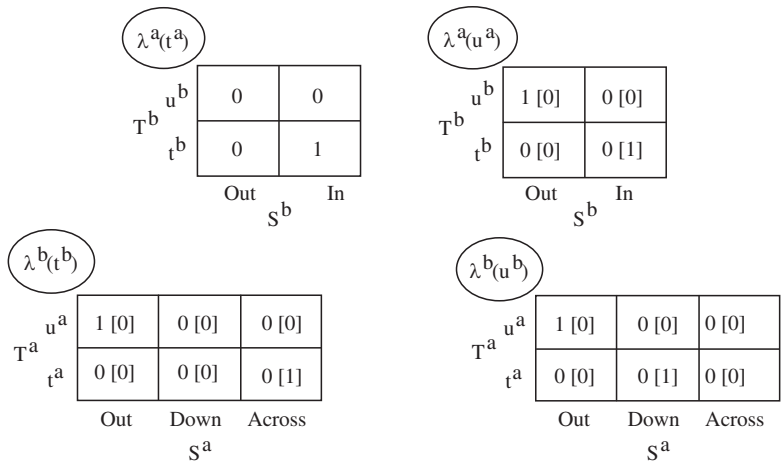
There are two types t^a, u^a for Ann. Type t^a for Ann has the measure shown in the top-left matrix. It assigns probability 1 to (In, t^b) for Bob. Type u^a has two associated measures – shown in the top-right matrix. The first measure (the numbers without parentheses) assigns probability 1 to (Out, u^b) for Bob. In this case, we also specify a second measure for Ann, because we want to specify what Ann thinks at her second node, too. Reaching this node is assigned positive probability (in fact, probability 1) under Ann's type t^a, but probability 0 under her type u^a. So, for type u^a, there isn't a well-defined conditional probability measure at Ann's second node. This is why we (separately) specify a second measure for Ann's type u^a: it is the measure in square brackets. If type u^a, Ann assigns probability 1 to (In, t^b) at her second node.

There are also two types t^b, u^b for Bob. Both types initially assign probability 1 to Ann's playing Out. For both of Bob's types, there isn't a well-defined conditional probability measure at his node. At his node, Bob's type t^b assigns probability 1 to $\{(Across, t^a)\}$, while his type u^b assigns probability 1 to $\{(Down, t^a)\}$.

This is a simple illustration of the concept of a *conditional probability system* (*CPS*), due to Rényi (1955). A CPS specifies a family of conditioning events E and a measure p_E for each such event, together with certain restrictions on these measures. The interpretation is that p_E is what the player believes, after observing E. Even if $p_\Omega(E) = 0$ (where Ω is the entire space), the measure p_E is still specified. That is, even if E is 'unexpected', the player has a measure if E nevertheless happens. This is why CPS's are well-suited to epistemic analysis of game trees – where we need to be able to describe how players react to the unexpected.

Myerson (1991, ch. 1) provided a preference-based axiomatization of a class of CPS's. Battigalli and Siniscalchi (1999, 2002) further developed both the pure theory and the game-theoretic application of CPS's (see below).

Suppose the true state in Fig. 3.2 is (*Down*, t^a, *In*, t^b). In particular, Ann plays *Down*, expecting Bob to play *In*. Bob plays *In*, expecting (at his node) Ann to play *Across*. Ann expects a payoff of 4 (and gets this). Bob expects a payoff of 6 (but gets only 3). In everyday language, we can say that Ann successfully bluffs Bob. (At the state



Epistemic Game Theory: Complete Information, Fig. 3.2

(*Down*, t^a, *In*, t^b), the bluff works. By contrast, at the state (*Down*, t^a, *Out*, u^b), Ann attempts the bluff and it fails.)

But what about epistemic conditions? Are the players rational in this situation? Does each think the other is rational? And so on.

To answer, we need a definition of rationality with CPS's. Fix a strategy-type pair (s^i, t^i), where t^i is associated with a CPS. Call this pair **rational (in the tree)** if the following holds: fix any information set H for i allowed by s^i, and look at the measure on the other players' strategies, given H. (This means given the event that the other players' strategies allow H.) Require that s^i maximizes i's expected payoff under this measure, among all strategies r^i of i that allow H.

With this definition, the rational strategy-type pairs in Fig. 3.2 are (*Down*, t^a), (*Out*, u^a), (*In*, t^b), and (*Out*, u^b).

Next, what does Ann think about Bob's rationality? To answer, we need a CPS-analogue to belief (as defined in section "Early Results"). Ben Porath (1997) proposed the following (we have taken the liberty of changing terminology, for consistency with 'strong belief' below): Say player i **initially believes** event E if, under i's CPS, E gets probability 1 at the root of the tree. (Formally, the conditioning event consists of all strategy profiles of the other players.) Battigalli and Siniscalchi (2002) strengthened this definition to: Say player i **strongly believes** event E if, under i's CPS, E gets probability 1 at every information set at which E is possible. Under initial belief, E also gets probability 1 at any information set H that gets positive probability under i's initial measure (that is, i's measure given the root). This is just standard conditioning on non-null events. But under strong belief, this conclusion holds for any information set H which has a non-empty intersection with E – even if H is null under i's initial measure. This is why strong belief is stronger than initial belief.

Let us apply these definitions to Fig. 3.2. Does Ann initially believe that Bob is rational? Yes. Both of Ann's types initially believe Bob is rational. Type t^a initially assigns probability 1 to the rational pair (*In*, t^b). Type u^a initially assigns probability 1 to the rational pair (*Out*, u^b). In fact, both types strongly believe Bob is rational. Since, under type t^a, Ann's second node gets positive probability (in fact, probability 1) under her initial measure, we need only check this for type u^a. But at Ann's second node, type u^a assigns probability 1 to the rational pair (*In*, t^b).

Turning to Bob, both of his types initially believe that Ann is rational. Type u^b even strongly believes Ann is rational; but type t^b doesn't. This is because, at Bob's node, type t^b assigns positive probability (in fact, probability 1) to the irrational pair (*Across*, t^a).

Staying with initial belief (we come back to strong belief below), we can parallel Definition 2.1 and define inductively **rationality and mth-order initial belief of rationality (RmIBR)** at a state of a type structure, and **rationality and common initial belief of rationality (RCIBR)** (see Ben Porath, 1997). In Fig. 3.2, since all four types initially believe the other player is rational, a simple induction gives that at the state (*Down*, t^a, *In*, t^b) for instance, RCIBR holds.

In words, Ann plays across at her first node, believing (initially) that Bob will play *In*, so she can get a payoff of 4. Why would Bob play In? Because he initially believes that Ann plays *Out*. But in the probability-0 event that Ann plays across at her first node, Bob then assigns probability 1 to Ann's playing across at her second node – that is, to Ann's being irrational. He therefore (rationally) plays *In*. All this is consistent with RCIBR.

Conditions for Backward Induction

Interestingly, this is exactly the line of reasoning which, as we said, was the original stimulus for investigating the foundations of BI. So, there is no difficulty with it – we've just seen a formal set-up in which it holds. The resolution of the BI puzzle is simply to accept that the BI path may not result.

But one can also argue that RCIBR is not the right condition: it is too weak. In the above example, Bob realizes that he might be 'surprised' in the play of the game – that's why he has a CPS, not just an ordinary probability measure. If he realizes he might be surprised, should he abandon his

(initial) belief that Ann is rational when he is surprised? Bob's type t^b does so. This is the step taken by Battigalli and Siniscalchi (2002) with their concept of strong belief. The argument says that we want t^b to strongly believe, not just initially believe, that Ann is rational. Type t^b will strongly believe Ann is rational if we move the probability-1 weight (in square brackets) on (*Across*, t^a) to (*Down*, t^a). But now (*In*, t^b) isn't rational for Bob, so Ann doesn't (even initially) believe Bob is rational. It looks as if the example unravels.

We can again parallel Definition 2.1 and define inductively **rationality and *m*th-order strong belief of rationality (R*m*SBR)**, and **rationality and common strong belief of rationality (RCSBR)** (see Battigalli and Siniscalchi, 2002). The question is then: does RCSBR yield BI?

The answer is yes. *Fix a CPS-based type structure for n-legged Centipede* (Fig. 4.1), *and a state at which there is RCSBR. Then Ann plays Out*. The result follows from Friedenberg (2002), who shows that in a PI game (satisfying certain payoff restrictions), RCSBR yields a Nash-equilibrium outcome. In Centipede, there is a unique Nash path and it coincides with the BI path. Of course, this isn't true in general.

Example 4.1 (A Second Coordination Game) *Consider the coordination game in* Fig. 4.2 *and the associated CPS-based type structure in* Fig. 4.3.

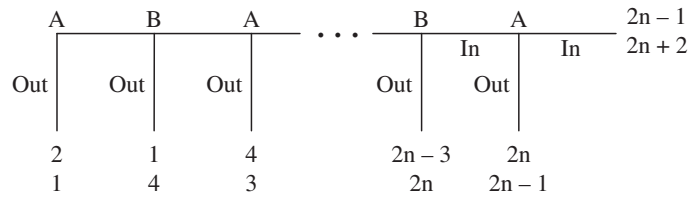
The rational strategy-type pairs are (*Out,* t^a) *and* (*Out,* t^b) *for Ann and Bob respectively. Ann's type* t^a *strongly believes* {(*Out,* t^b)}, *and Bob's type* t^b *strongly believes* {(*Out,* t^a)}. *By induction, RCSBR holds at the state* (*Out,* t^a, *Out,* t^b).

Here, the BI path need not be played under RCSBR. The key is to see that both (*Down*, t^a) and (*Across*, t^a) are irrational for Ann, since she (strongly) believes Bob plays *Out*. So at his node, Bob can't believe Ann is rational. If he considers it sufficiently more likely Ann will play *Down* rather than *Across*, he will rationally play *Out* (as happens). In short, if Ann doesn't play *Out*, she is irrational and so 'all bets are off' as to what she will do. She could play *Down*.

This situation may be surprising, at least at first blush, but there does not appear to be anything conceptually wrong with it. Indeed, it points to an interesting way in which the players in a game can literally be trapped by their beliefs – which here prevent them from getting their mutually preferred (3, 3) outcome.

But one can also argue differently. If Ann forgoes the payoff of 2 she can get by playing *Out* at the first node, then surely she must be playing *Across* to get 3. Playing *Down* to get 0 makes little sense since this is lower than the payoff she gave up at the first node. (This is forward-induction reasoning à la Kohlberg and Mertens, 1986, Section 2.3, introduced in the context of non-PI games. Interestingly, epistemic analysis makes clear that the issue already arises in PI games, such as Fig. 4.2.) But if Bob considers *Across* (sufficiently) more likely than *Down*, he will play *In*. Presumably then, Ann will indeed play *Across*, and the BI path results.

There is no contradiction with the previous analysis because in Fig. 4.3 Ann is irrational once she doesn't play *Out*, so we can't say Ann should then rationally play *Across* not *Down*. To make *Across* rational for Ann, we have to add more types to the structure – specifically, we would want to add a second type for Ann that assigns (initial) probability 1 to Bob's playing *In* not *Out*. This key insight is due to Stalnaker (1998) and Battigalli and Siniscalchi (2002).

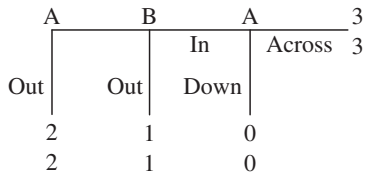


Epistemic Game Theory: Complete Information, Fig. 4.1

Battigalli and Siniscalchi formulate a general result of this kind. They consider a **complete** CPS-based type structure, which contains, in a certain sense, every possible type for each player (a complete type structure will be uncountably infinite), and prove: *Fix a complete CPS-based type structure. If there is RCSBR at the state* $(s^1, t^1, ..., s^n, t^n)$, *then the strategy profile* $(s^1, ..., s^n)$ *is extensive-form rationalizable. Conversely, if the profile* $(s^1, ..., s^n)$ *is extensive-form rationalizable, then there is a state* $(s^1, t^1, ..., s^n, t^n)$ *at which there is RCBR.*

The extensive-form rationalizability strategies (Pearce, 1984) yield the BI outcome in a PI game (under an assumption ruling out certain payoff ties; Battigalli, 1997), so the Battigalli and Siniscalchi analysis gives epistemic conditions for BI.

There are other routes to getting BI in PI games. Asheim (2001) develops an epistemic analysis using the properness concept (Myerson, 1978). Go back to Example 4.1. The properness idea says that Bob's type t^b should view (*Across*, t^a) as infinitely more likely than (*Down*, t^a) since *Across* is the less costly 'mistake' for Ann, given her type t^a. Unlike the completeness route taken above, the irrationality of both *Down* and *Across* (given Ann's type t^a) is accepted. But the relative ranking of these 'mistakes' must be in the right order. With this ranking, Bob is irrational to play *Out* rather than *In*. Ann presumably will play *Across*, and we get BI again. Asheim (2001) formulates a general such result.

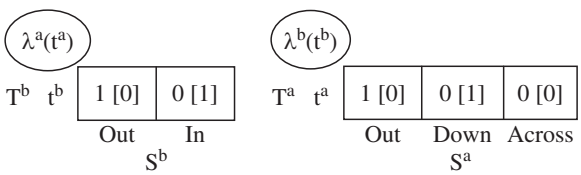


Epistemic Game Theory: Complete Information, Fig. 4.2

Another strand of the literature on BI employs knowledge models rather than belief models. As pointed out in Example 1.1, players' beliefs don't have to be correct in any sense. For example, a type might even assign probability 1 to a strategy-type pair for another player different from the actual one. Knowledge as usually formalized is different, in that if a player knows an event E, then E indeed happens.

Aumann (1995) formulates a knowledge-based epistemic model for PI trees. In his set-up, the condition of common knowledge of rationality implies that the players choose their BI strategies. Stalnaker (1996) finds that non-BI outcomes are possible, under a different formulation of the same condition. The explanation lies in differences in how counterfactuals are treated. These play an important role in a knowledge-based analysis, when we talk about what a player thinks at an information set that cannot be reached given what he knows. Halpern (2001) provides a synthesis in which these differences can be understood. See also the exchange between Binmore (1996) and Aumann (1996), and the analyses by Samet (1996), Balkenborg and Winter (1997), and Halpern (1999).

Aumann (1998) provides knowledge-based epistemic conditions under which Ann plays *Out* in Centipede. The conditions are weaker than in his (1995) paper, and the conclusion weaker (about outcomes not strategies). There is an obvious parallel between this result and the belief-based result on Centipede we stated above (also about outcomes). More generally, there may be an analogy between counterfactuals in knowledge models and extended probabilities in belief models. But, for one thing, completeness is crucial to the belief-based approach, as we have seen, and an analogous concept does not appear



Epistemic Game Theory: Complete Information, Fig. 4.3

to be present in the knowledge-based approach. As yet, there does not appear to be any formal treatment of the relationship between the two approaches.

Next Steps: Weak Dominance

Extending the epistemic analysis of games from the matrix to the tree has been the focus of much recent work in the literature. Another area has been extending the analysis on the matrix from strong dominance (described in section "Early Results") to weak dominance.

Weak dominance (admissibility) says that a player considers as possible (even if unlikely) any of the strategies for the other players. In the game context, we are naturally led to consider *iterated admissibility* (*IA*) – the weak-dominance analogue to IU. This is an old concept in game theory, going back at least to Gale (1953). Like BI, it is a powerful solution concept, delivering sharp answers in many games – Bertrand, auctions, voting games, and others. (Mertens, 1989, p. 582, and Marx and Swinkels, 1997, pp. 224–5, list various games involving weak dominance.)

But, also like BI, there is a conceptual puzzle. Suppose Ann conforms to the admissibility requirement, so that she considers possible any of Bob's strategies. Suppose Bob also conforms to the requirement, and this leads him not to play a strategy, say *L*. If Ann thinks Bob adheres to the requirement (as he does), then she can rule out Bob's playing *L*. But this conflicts with the requirement that she not rule anything out (see Samuelson, 1992).

Can a sound argument be made for IA? To investigate this, the epistemic tools of section "Epistemic Analysis" have to be extended again.

Example 5.1 (Bertrand) Figure 5.1 *is a Bertrand pricing game, where each firm chooses a price in* {*0, 1, 2, 3*}. (*Ken Corts kindly provided this example*.) *The left payoff is to A, the right payoff to B. Each firm has capacity of two units and zero cost. Two units are demanded. If the firms charge the same price, they each sell one unit*. Figure 5.2 *is an associated type structure* (*with one type for each player*).

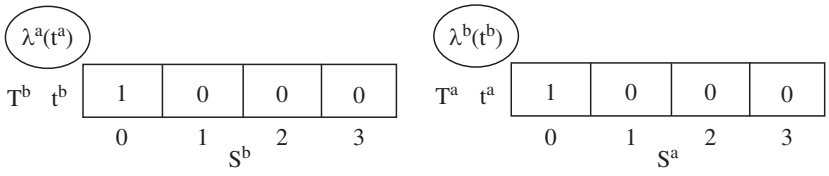
A \ B	3	2	1	0
3	3, 3	0, 4	0, 2	0, 0
2	4, 0	2, 2	0, 2	0, 0
1	2, 0	2, 0	1, 1	0, 0
0	0, 0	0, 0	0, 0	0, 0

Epistemic Game Theory: Complete Information, Fig. 5.1

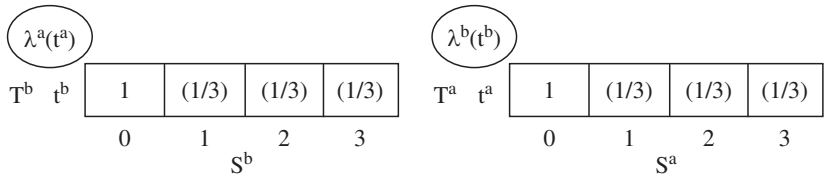
The rational strategy-type pairs are $R_1^a = \{0,1,2,3\} \times \{t^a\}$ *and* $R_1^b = \{0,1,2,3\} \times \{t^b\}$. *Since both types assign positive probability only to a rational strategy-type pair for the other player, we get* $R_m^a = R_1^a$ *and* $R_m^b = R_1^b$ *for all m. In particular, there is RCBR at the state* (*3,* t^a*, 3,* t^b).

But a price of 3 is inadmissible (as is a price of 0). The IA set is just {(1,1)}, where each firm charges the lowest price above cost. (This is a plausible scenario: while pricing at cost is inadmissible, competition forces price down to the first price above cost.)

A tool to incorporate admissibility is *lexicographic probability systems* (*LPS*'s), introduced and axiomatized by Blume et al. (1991a, b). An LPS specifies a sequence of probability measures. The interpretation is that the first measure is the player's primary hypothesis about the true state. But the player recognizes that his primary hypothesis might be mistaken, and so also forms a secondary hypothesis. This is his second measure. Then his tertiary hypothesis, and so on. The primary states can be thought of as infinitely more likely than the secondary states, which are infinitely more likely than the tertiary states, and so on. Stahl (1995), Stalnaker (1998), Asheim (2001), Brandenburger et al. (2006), and Asheim and Perea (2005), among other papers, use LPS's.



Epistemic Game Theory: Complete Information, Fig. 5.2



Epistemic Game Theory: Complete Information, Fig. 5.3

Example 5.2 (Bertrand Contd) Figure 5.3 *is a type structure for Bertrand* (Fig. 5.1) *that now specifies LPS's.*

Each player has a primary hypothesis which assigns probability 1 to the other player's charging a price of 0. But each player also has a secondary hypothesis that assigns equal probability to each of the three remaining choices for the other player. This measure is shown in parentheses. Note that every state (that is, strategy-type pair) gets positive probability under some measure. But states can also be ruled out, in the sense that they can be give infinitely less weight than other states.

What about epistemic conditions? Are the players rational in this situation? Does each think the other is rational? And so on.

To answer, we need a definition of rationality with LPS's. Fix strategy-type pairs (s^i, t^i) and (r^i, t^i) for player i, where t^i is now associated with an LPS. Calculate the tuple of expected payoffs to i from s^i, using first the primary measure associated with t^i, then the secondary measure associated with t^i, and so on. Calculate the corresponding tuple for r^i. If the first tuple lexicographically exceeds the second, then s^i is preferred to r^i. (If $x = (x_1,..., x_n)$ and $y = (y_1,..., y_n)$, then x lexicographically exceeds y if $y_j > x_j$ implies $x_k > y_k$ for some $k < j$.) A strategy-type pair (s^i, t^i) is **rational (in the lexicographic sense)** if s^i is maximal under this ranking.

So $(3, t^a)$ and $(3, t^b)$ are irrational. All choices give each player an expected payoff of 0 under the primary measure. But a price of 2 gives each player an expected payoff of 2 under the secondary measure, as opposed to an expected payoff of 1 from a price of 3. Conceptually, we want $(3, t^a)$ and $(3, t^b)$ to be irrational (because a price of 3 is inadmissible).

What does each player think about the other's rationality? For this, we again need an LPS-based definition. An early candidate in the literature was: Say player i **believes** event E **at the 1st level** if E gets primary probability 1 under i's LPS (Börgers, 1994; Brandenburger, 1992). A stronger concept is: Say i **assumes** E if all states not in E are infinitely less likely than all states in E, under i's LPS (Brandenburger, Friedenberg and Keisler, 2006). In other words, a player who assumes E recognizes E may not happen, but is prepared to 'count on' E versus not-E.

In Fig. 5.3, type t^a doesn't 1st-level believe (so certainly doesn't assume) the other player is rational. Likewise with t^b. Again, this is right conceptually.

Conditions for Iterated Admissibility

Once again we can parallel Definition 2.1 and define inductively **rationality and *m*th-order 1st-level belief of rationality (R*m*1BR)** at a state of a type structure, and **rationality and common 1st-level belief of rationality (RC1BR)**. Likewise, one can define **rationality and *m*th-order assumption of rationality (R*m*AR), and rationality and common assumption of rationality (RCAR).** What do these conditions yield?

In fact, just as we saw in sections "Next Steps: The Tree and Conditions for Backward Induction" that neither RCIBR not RCSBR yields BI, so neither RC1BR nor RCAR yields IA. RC1BR is characterized by the $S^{\infty}W$ concept (Dekel and Fudenberg, 1990), that is, the set of strategies that remain after one round of deletion of inadmissible strategies followed by iterated deletion of strongly dominated strategies. RCAR is characterized by the self-admissible set concept (Brandenburger, Friedenberg and Keisler, 2006). Self-admissible sets may be viewed as the weak-dominance analogue to Pearce (1984) best-response sets.

But while the IA set is one self-admissible set in a game, there may well be others. To select the IA set, a completeness assumption is needed, similar to section "Conditions for Backward Induction": *Fix a complete LPS-based type structure. If there is RmAR at the state* $(s^1, t^1, ..., s^n, t^n)$*, then the strategy profile* $(s^1, ..., s^n)$ *survives* $(m + 1)$ *rounds of iterated admissibility. Conversely, if the profile* $(s^1, ..., s^n)$ *survives* $(m + 1)$ *rounds of iterated admissibility, then there is a state* $(s^1, t^1, ..., s^n, t^n)$ *at which there is RmAR* (Brandenburger, Friedenberg and Keisler, 2006).

This result is stated for R*m*AR and not RCAR. See the next section for the reason. Of course, for a given game, there is an m such that IA stabilizes after m rounds.

IA yields the BI outcome in a PI game (again ruling out certain payoff ties; Marx and Swinkels, 1997), so, understanding IA gives, in particular, another analysis of BI.

Related analyses of IA include Stahl (1995) and Ewerhart (2002). Stahl uses LPS's and directly assumes that Ann considers one of Bob's strategies infinitely less likely than another if the first is eliminated on an earlier round of IA than the second. Ewerhart gives an analysis of IA couched in terms of provability (from mathematical logic).

Strategic Versus Extensive Analysis

Kohlberg and Mertens (1986, Section 2.4) argued that a 'fully rational' analysis of games should be invariant – that is, should depend only on the fully reduced strategic form of a game. (This is the strategic form after elimination of any – pure – strategies that are duplicates or convex combinations of other strategies.) In this, they appealed to early results in game theory (Dalkey, 1953; Thompson, 1952) which established that two trees sharing the same reduced strategic form differ from each other by a (finite) sequence of elementary transformations of the tree, each of which can be argued to be 'strategically inessential'. Kohlberg and Mertens added a fourth transformation involving convex combinations, to get to the fully reduced strategic form.

In decision theory, invariance is implied by (and implies) admissibility. (Kohlberg and Mertens, 1986, Section 2.7, gave the essential idea. See Brandenburger, 2007, for the decision-theory argument.) If we build up our game analysis using a decision theory that satisfies admissibility, we can hope to get invariance at this level too. LPS-based decision theory satisfies admissibility. Indeed, IA, and also the $S^{\infty}W$ and self-admissible set concepts, are invariant in the Kohlberg–Mertens sense. The extensive-form rationalizability concept (section "Conditions for Backward Induction") is not.

There does appear to be a price paid for invariance, however. The extensive-form conditions of RCSBR and (CPS-based) completeness are consistent (in any tree). That is, for any tree, we can build a complete type structure and find a state at which RCSBR holds. But Brandenburger et al. (2006) show the strategic-form conditions of RCAR and (LPS-based) completeness are inconsistent (in any matrix satisfying a non-triviality condition).

A possible interpretation is that rationality, even as a theoretical concept, appears to be inherently limited. There are purely theoretical limits to the Kohlberg-Mertens notion of a 'fully rational' analysis of games.

The epistemic programme has uncovered a number of impossibility results (see epistemic game theory: beliefs and types for some others). We don't see this as a deficiency of the programme, but rather as a sign it has reached a certain depth and maturity. Also, central to the programme is the analysis of scenarios (we have seen several in this survey) that are 'a long way from' these theoretical limits. Under the epistemic approach to game theory there is not one right set of assumptions to make about a game.

See Also

- ▶ Epistemic Game Theory: An Overview
- ▶ Epistemic Game Theory: Beliefs and Types
- ▶ Epistemic Game Theory: Incomplete Information
- ▶ Game Theory
- ▶ Nash Equilibrium, Refinements of

This survey is based on Brandenburger (2007). I am grateful to Springer for permission to use this material. I owe a great deal to joint work and many conversations with Robert Aumann, Eddie Dekel, Amanda Friedenberg, Jerry Keisler and Harborne Stuart. My thanks to Konrad Grabiszewski for important input, John Nachbar for very important editorial advice, and Michael James for valuable assistance. The Stern School of Business provided financial support.

Bibliography

Asheim, G. 2001. Proper rationalizability in lexicographic beliefs. *International Journal of Game Theory* 30: 453–478.

Asheim, G., and A. Perea. 2005. Sequential and quasi-perfect rationalizability in extensive games. *Games and Economic Behavior* 53: 15–42.

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67–96.

Aumann, R. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18.

Aumann, R. 1995. Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8: 6–19.

Aumann, R. 1996. Reply to Binmore. *Games and Economic Behavior* 17: 138–146.

Aumann, R. 1998a. On the centipede game. *Games and Economic Behavior* 23: 97–105.

Aumann, R. 1998b. Common priors: A reply to Gul. *Econometrica* 66: 929–938.

Aumann, R., and A. Brandenburger. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63: 1161–1180.

Balkenborg, D., and E. Winter. 1997. A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics* 27: 325–345.

Basu, K. 1990. On the existence of a rationality definition for extensive games. *International Journal of Game Theory* 19: 33–44.

Battigalli, P. 1997. On rationalizability in extensive games. *Journal of Economic Theory* 74: 40–61.

Battigalli, P., and M. Siniscalchi. 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory* 88: 188–230.

Battigalli, P., and M. Siniscalchi. 2002. Strong belief and forward-induction reasoning. *Journal of Economic Theory* 106: 356–391.

Ben Porath, E. 1997. Rationality, Nash equilibrium, and backward induction in perfect information games. *Review of Economic Studies* 64: 23–46.

Bernheim, D. 1984. Rationalizable strategic behavior. *Econometrica* 52: 1007–1028.

Bicchieri, C. 1988. Strategic behavior and counterfactuals. *Synthese* 76: 135–169.

Bicchieri, C. 1989. Self-refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis* 30: 69–85.

Binmore, K. 1987. Modelling rational players I. *Economics and Philosophy* 3: 179–214.

Binmore, K. 1996. A note on backward induction. *Games and Economic Behavior* 17: 135–137.

Blume, L., A. Brandenburger, and E. Dekel. 1991a. Lexicographic probabilities and choice under uncertainty. *Econometrica* 59: 61–79.

Blume, L., A. Brandenburger, and E. Dekel. 1991b. Lexicographic probabilities and equilibrium refinements. *Econometrica* 59: 81–98.

Bonanno, G. 1991. The logic of rational play in games of perfect information. *Economics and Philosophy* 7: 37–65.

Bonanno, G., and K. Nehring. 1999. How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory* 28: 409–434.

Börgers, T. 1994. Weak dominance and approximate common knowledge. *Journal of Economic Theory* 64: 265–276.

Brandenburger, A. 1992. Lexicographic probabilities and iterated admissibility. In *Economic analysis of markets and games*, ed. P. Dasgupta, D. Gale, O. Hart, and E. Maskin. Cambridge, MA: MIT Press.

Brandenburger, A. 2007. The power of paradox: Some recent results in interactive epistemology. *International Journal of Game Theory* 35: 465–492.

Brandenburger, A., and E. Dekel. 1987. Rationalizability and correlated equilibria. *Econometrica* 55: 1391–1402.

Brandenburger, A., Friedenberg, A. and Keisler, H.J. 2006. Admissibility in games. Unpublished, Stern School of Business, New York University.

Dalkey, N. 1953. Equivalence of information patterns and essentially determinate games. In *Contributions to the theory of games*, ed. H. Kuhn and A. Tucker, Vol. 2. Princeton: Princeton University Press.

Dekel, E., and D. Fudenberg. 1990. Rational behavior with payoff uncertainty. *Journal of Economic Theory* 52: 243–267.

Ewerhart, C. 2002. Ex-ante justifiable behavior, common knowledge, and iterated admissibility. Unpublished, Department of Economics, University of Bonn.

Feinberg, Y. 2000. Characterizing common priors in terms of posteriors. *Journal of Economic Theory* 91: 127–179.

Friedenberg, A. 2002. When common belief is correct belief. Unpublished, Olin School of Business, Washington University.

Gale, D. 1953. A theory of *n*-person games with perfect information. *Proceedings of the National Academy of Sciences* 39: 496–501.

Gul, F. 1998. A comment on Aumann's Bayesian view. *Econometrica* 66: 923–927.

Halpern, J. 1999. Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory* 28: 315–330.

Halpern, J. 2001. Substantive rationality and backward induction. *Games and Economic Behavior* 37: 425–435.

Halpern, J. 2002. Characterizing the common prior assumption. *Journal of Economic Theory* 106: 316–355.

Harsanyi, J. 1967–8. Games with incomplete information played by 'Bayesian' players, I–III. *Management Science* 14: 159–182, 320–334, 486–502.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: A new rationale for mixed strategy equilibrium points. *International Journal of Game Theory* 2: 1–23.

Kohlberg, E., and J.-F. Mertens. 1986. On the strategic stability of equilibria. *Econometrica* 54: 1003–1037.

Marx, L., and J. Swinkels. 1997. Order independence for iterated weak dominance. *Games and Economic Behavior* 18: 219–245.

Mertens, J.-F. 1989. Stable equilibria – A reformulation. *Mathematics of Operations Research* 14: 575–625.

Morris, S. 1994. Trade with heterogeneous prior beliefs and asymmetric information. *Econometrica* 62: 1327–1347.

Myerson, R. 1978. Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 1: 73–80.

Myerson, R. 1991. *Game theory*. Cambridge, MA: Harvard University Press.

Pearce, D. 1984. Rational strategic behavior and the problem of perfection. *Econometrica* 52: 1029–1050.

Reny, P. 1992. Rationality in extensive form games. *Journal of Economic Perspectives* 6(4): 103–118.

Rényi, A. 1955. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae* 6: 285–335.

Rosenthal, R. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* 25: 92–100.

Samet, D. 1996. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior* 17: 230–251.

Samet, D. 1998a. Common priors and the separation of convex sets. *Games and Economic Behavior* 24: 172–174.

Samet, D. 1998b. Iterated expectations and common priors. *Games and Economic Behavior* 24: 131–141.

Samuelson, L. 1992. Dominated strategies and common knowledge. *Games and Economic Behavior* 4: 284–313.

Stahl, D. 1995. Lexicographic rationalizability and iterated admissibility. *Economic Letters* 47: 155–159.

Stalnaker, R. 1996. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy* 12: 133–163.

Stalnaker, R. 1998. Belief revision in games: Forward and backward induction. *Mathematical Social Sciences* 36: 31–56.

Tan, T., and S. Werlang. 1988. The Bayesian foundations of solution concepts of games. *Journal of Economic Theory* 45: 370–391.

Thompson, F. 1952. Equivalence of games in extensive form. Research Memorandum RM-759. The RAND Corporation.

Epistemic Game Theory: Incomplete Information

Aviad Heifetz

Abstract

In a game of incomplete information some of the players possess private information which may be relevant to the strategic interaction. Private information is modelled by a *type*

space, in which every type of each player is associated with a belief about the basic issues of uncertainty (like payoffs) and about the other players' types. At a *Bayesian equilibrium* each type chooses a strategy which maximizes its expected payoff given the choice of strategies by the other players' types. Bayesian equilibrium payoffs are often inefficient relative to the equilibrium payoffs that would result had the players been fully informed.

Keywords

Bayesian equilibrium; Bayesian strategies; Common knowledge; Epistemic game theory: incomplete information; Games with incomplete information; Private information

JEL Classifications

C7

A game of incomplete information is a game in which at least some of the players possess private information which may be relevant to the strategic interaction. The private information of a player may be about the payoff functions in the game, as well as about some exogenous, payoff-irrelevant events. The player may also form beliefs about other players' beliefs about payoffs and exogenous events, about their beliefs about the beliefs of others, and so forth.

Harsanyi (1967–8) introduced the idea that such a state of affairs can be succinctly described by a *type space*. With this formulation, T_i denotes the set of player i's *types*. Each type $t_i \in T_i$ is associated with a belief $\lambda_i(t_i) \in \Delta(K \times T_{-i})$ about some basic space of uncertainty, K, and the combination T_{-i} of the other players' types. The basic space of uncertainty K is called the space of *states of nature*, and $\Omega = K \times \Pi_{i \in I} T_i$, where I is the set of players, is called the space of *states of the world*.

A type space models a game of incomplete information once each state of nature $k \in K$ is associated with a payoff matrix of the game, or, more generally, with a payoff function u_i^k for each player $i \in I$. This payoff function specifies the player's payoff $u_i^k(s)$ for each combination of strategies $s = (s_i)_{i \in I} \in S = \Pi_{i \in I} S_i$ of the players. (In the particular case in which k is associated with a payoff matrix, that is, the game is such that each player has finitely many strategies, the payoffs $u_i^k(s)$ to the players $i \in I$ appear in the entry of the matrix corresponding to the combination of strategies $s = (s_i)_{i \in I}$.) As usual, the set of strategies S_i of player $i \in I$ may be a complex object by itself. For instance, it may be the set of mixed strategies over some set of pure strategies S_i^0. The payoff function of player i in the state of nature k is $u_i^k : S \to \mathbb{R}$.

Obviously, different types of a player may want to choose different strategies. Thus, a *Bayesian strategy* of player i in a game of incomplete information specifies the strategy $\sigma_i(t_i) \in S_i$ that the player chooses given each one of her types $t_i \in T_i$.

Given a profile of Bayesian strategies $\sigma = (\sigma_j : T_j \to S_j)_{j \in I}$ of the players, the expected payoff of player i of type t_i is

$$U_i(\sigma, t_i) = \sum_{(k, t_{-i}) \in K \times T_{-i}} u_i^k\left(\sigma_i(t_i), \sigma_{-i}(t)\right) \times \lambda_i(t_i)(k, t_{-i})$$

where $\sigma_{-i}(t_{-i}) = (\sigma_j(t_j))_{j \neq i}$. If there is a continuum of states of nature and types, the sum becomes an integral:

$$U_i(\sigma, t_i) = \int_{K \times T_{-i}} u_i^k\left(\sigma_i(t_i), \sigma_{-i}(t_{-i})\right) d\lambda_i(t_i)(k, t_{-i})$$

(In this case, the expected payoff function $U_i(\sigma, t_i)$ is well defined if the Bayesian strategies $\sigma_j : T_j \to S_j$ are measurable functions and if the payoff function $u_i : K \times S \to R$ is measurable as well; we omit the details of this technical requirement).

We assume that the players are expected payoff maximizers. Thus, player i prefers the Bayesian strategy σ over σ' if and only if $U_i(\sigma, t_i) \geq U_i(\sigma', t_i)$ for each of her types $t_i \in T_i$. It follows that given a Bayesian strategy profile σ_{-i} of the other players, the Bayesian strategy σ_i is a *best reply* of player i if for any other strategy σ' of hers, $U_i((\sigma_i, \sigma_{-i}), t_i) \geq U_i((\sigma', \sigma_{-i}), t_i)$ for each of her

types $t_i \in T_i$. A *Bayes–Nash equilibrium* or a *Bayesian equilibrium* is a profile of Bayesian strategies $\sigma^* = (\sigma_{i*})_{i \in I}$ such that σ_{i*} is a best reply against σ_{-i*} for every player $i \in I$.

A simple, discrete variant of an example by Gale (1996) may clarify these abstract definitions. There are two investors $i = 1, 2$ and three possible states of nature $k \in K = \{-1, 0, 1\}$. Each investor i only knows her own type

$$t_i \in T_i = \{-10, -6, -2, 2, 6, 10\}.$$

Every type t_i of investor i believes that all of the other investor's types $t_j \in T_j, j \neq i$, are equally likely, so that each of them has probability $\frac{1}{6}$. Moreover, every type t_i believes that the state of nature is $k = 1$ when $t_i + t_j > 0$; that the state of nature is $k = 0$ when $t_i + t_j = 0$; and that the state of nature is $k = -1$ when $t_i + t_j < 0$. Formally, the belief $\lambda_i(t_i)$ of type $t_i \in T_i$ is defined by

$$\lambda_i(t_i)(k, t_j) = \begin{cases} \frac{1}{6} & k \text{ has the same sign as } t_i + t_j \\ 0 & \text{otherwise} \end{cases}$$

The investors cannot communicate their types to one another. They can invest in at most one of two available investment periods. Each investor has three relevant strategies: invest *immediately*, in the first period; *wait* to the second period and invest only if the other investor has invested in the first period; or *never* invest. The payoff of each of the investors depends on the state of nature $k \in K = \{-1, 0, 1\}$ and on her own investment strategy, but not on the investment strategy of the other investor. The payoffs are as follows:

- Investing *immediately* when the state of nature is k yields investor i a payoff of k

$$u_i^k(\text{`}immediately\text{'}, \cdot) = k$$

(The $\cdot$ stands for the investment decision of the other investor $j \neq i$, which, as we said, does not effect the payoff of investor i.)

- If investor i chooses to *wait* to the second period and invest only if the other investor has invested in the first period, investor i's payoff in the state of nature k is

$$u_i^k(\text{`}wait\text{'}, \cdot) = -\frac{3}{4}k.$$

- If the investor *never* invests, her payoff is 0 irrespective of the state of nature:

$$u_i^k(\text{`}never\text{'}, \cdot) = 0.$$

How will the different types behave at a Bayesian equilibrium? The type $t_i = 10$ assesses that by investing immediately her expected payoff is

$$Ui(\text{`}immediately\text{'}, 10) = \frac{1}{6} \times 0 + \frac{5}{6} \times 1 = \frac{5}{6}$$

(immediate investment yields 0 in case $t_j = -10$, and yields 1 in case $t_j = -6, -2, 2, 6, 10$). This is higher than $\frac{3}{4}$, the maximum payoff she could possibly get by waiting for the second period, and higher than the payoff 0 of never investing. So at a Bayesian equilibrium

$$\sigma_i^*(10) = \text{`}immediately\text{'}, \quad i = 1, 2.$$

Next, the expected payoff to the type $t_i = 6$ from immediate investment is

$$Ui(\text{`}immediately\text{'}, 6) = \frac{1}{6} \times (-1) + \frac{1}{6} \times 0 + \frac{4}{6} \times 1 = \frac{1}{2}$$

(immediate investment yields 1 unless $t_j = -10$, in which case the payoff is -1, or $t_j = -6$, in which case the payoff is 0). So investing immediately is preferred for her over never investing. But how about waiting until the second period? That's an inferior option as well, since the types $t_j = -10, -6, -2$ will never invest in the first period (this would yield them a negative expected payoff). So only the positive types $t_j = 2, 6, 10$ could *conceivably* invest immediately, with overall probability reaching at most $\frac{3}{6}$. So waiting to see if they invest yields to the type $t_i = 6$ an expected payoff not higher $\frac{3}{6} \times \frac{3}{4} = \frac{3}{8}$, which is smaller than $\frac{1}{2}$. We conclude that the preferable strategy of $t_i = 6$ at equilibrium is

$$\sigma_i^*(6) = \text{`}immediately\text{'},\ \ i = 1, 2.$$

What about $t_i = 2$? Immediate investment yields her

$$Ui(\text{`}immediately\text{'}, 2) = \frac{2}{6} \times (-1) + \frac{1}{6} \times 0 + \frac{3}{6} \times 1 = \frac{1}{6}$$

(-1 is the payoff when $t_j = -10, -6$; 0 is the payoff when $t_j = -2$; the payoff is 1 otherwise). However, given that the types $t_j = 6,10$ invest immediately at equilibrium, and that the negative types $t_j = -10, -6, -2$ do not invest immediately, the type $t_i = 2$ figures out that by waiting and investing only if the other investor has invested first would yield her an expected payoff

$$Ui(\text{`}wait\text{'}, 2) \geq \frac{2}{6} \times \frac{3}{4} = \frac{1}{4} > \frac{1}{6}$$

($\frac{2}{6}$ is the probability assigned by $t_i = 2$ to the event that $t_j \in \{6, 10\}$ and hence j invests immediately, and $\frac{3}{4}$ is the payoff from the second period investment). Thepreferred strategy of $t_i = 2$ at equilibrium is therefore

$$\sigma_i^*(2) = \text{`}wait\text{'},\ \ i = 1, 2.$$

We can now compute inductively, in a similar way, that also

$$\sigma_i^*(-2) = \text{`}wait\text{'},\ \ i = 1, 2 \sigma_i^*(-6) = \text{`}wait\text{'},\ \ i = 1, 2$$

and that

$$\sigma_i^*(10) = \text{`}never\text{'},\ \ i = 1, 2.$$

Notice that the equilibrium in the example is inefficient. For instance, when the pair of types is $(t_1,t_2) = (2, 2)$ the investment is profitable, but both investors wait to see if the other one invests, and thus end up not investing at all. In this case, behaviour would become efficient if the investors could communicate their types to each other. Indeed, they would have been happy to do so, because their interests are aligned.

Obviously, there are other strategic situations with incomplete information in which the interests of the players are not completely aligned. For example, a potential seller of an object would like to strike a deal with a potential buyer at a price which is as high as possible, while the potential buyer would like the price to be as low as possible. That's why the traders might not volunteer to communicate honestly their private valuations of the object, even if they are technically able to do so. Still, in case the buyer values the object more than the seller, they would both prefer to trade at some price in-between their valuations rather than forgoing trade altogether. Therefore, the traders would nevertheless like to avoid a complete lack of communication. Myerson and Satterthwaite (1983) phrase general conditions under which no Bayesian equilibrium of any trade mechanism is ever fully efficient due to this tension between interests alignment and interests mismatch. Under these conditions, even if the traders are able to communicate their private information, at no Bayesian equilibrium does trade take place in all instances in which there exist gains from trade.

In the above variant of Gale's example we were able to find the unique Bayesian equilibrium using iterative dominance arguments. We have iteratively crossed out strategies that are inferior for some types, which enabled us to eliminate inferior strategies for other types, and so forth. As in games of complete information, this technique is not applicable in general, and there are games with incomplete information in which a Bayesian equilibrium is not the outcome of any process of iterative elimination of dominated strategies (Battigalli and Siniscalchi 2003; Dekel et al. 2007).

Games with incomplete information are discussed in many game theory textbooks (for example, Dutta 1999; Gibbons 1992; Myerson 1991; Osborne 2003; Rasmusen 1989; Watson 2002). Aumann and Heifetz (2002), Battigalli and Bonanno (1999) and Dekel and Gul (1997) are advanced surveys.

See Also

- ▶ Epistemic Game Theory: An Overview
- ▶ Epistemic Game Theory: Beliefs and Types
- ▶ Epistemic Game Theory: Complete Information
- ▶ Game Theory

Bibliography

Aumann, R.J., and A. Heifetz. 2002. Incomplete information. In *Handbook of game theory*, ed. R.J. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland.

Battigalli, P., and G. Bonanno. 1999. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics* 53: 149–225.

Battigalli, P., and M. Siniscalchi. 2003. Rationalization with incomplete information. *Advances in Theoretical Economics* 3(1), article 3. Online. Available at http://www.bepress.com/bejte/advances/vol3/iss1/art3. Accessed 25 Apr 2007.

Dekel, E., D. Fudenberg, and S. Morris. 2007. Interim correlated rationalizability. *Theoretical Economics* 2: 15–40.

Dekel, E., and F. Gul. 1997. Rationality and knowledge in game theory. In *Advances in economics and econometrics*, ed. D. Kreps and K. Wallis. Cambridge, UK: Cambridge University Press.

Dutta, P.K. 1999. *Strategies and games: Theory and practice*. Cambridge, MA: MIT Press.

Gale, D. 1996. What have we learned from social learning. *European Economic Review* 40: 617–628.

Gibbons, R. 1992. *Game theory for applied economists*. Princeton: Princeton University Press.

Harsanyi, J.C. 1967–8. Games with incomplete information played by Bayesian players, parts I–III. *Management Science* 14: 159–182, 320–334, 486–502.

Myerson, R. 1991. *Game theory: Analysis of conflict*. Cambridge, MA: Harvard University Press.

Myerson, R., and M. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.

Osborne, M. 2003. *Introduction to game theory.* Oxford: Oxford University Press.

Rasmusen, E. 1989. *Games and information: An introduction to game theory.* Oxford: Basil Blackwell.

Watson, J. 2002. *Strategy: An introduction to game theory.* New York: W.W. Norton.

Epistemological Issues in Economics

Shaun Hargreaves-Heap and Martin Hollis

Economics has raised such high hopes with its sophisticated techniques that the lack of agreed findings sets a puzzle. It will be suggested here that this lack of agreement reflects an epistemological puzzle, partly of how to recognize a causal law and partly of whether the search should be for causal laws in any case.

Start with the hypothetico-deductive method and the standard idea that economic theory advances when hypotheses confront the empirical evidence, as laid out in textbooks of positive economics, like Lipsey (1980). The backdrop is an empiricist picture of the natural world as an ordered realm, independent of our concepts, beliefs, hypotheses and conjectures about it. Science captures the order by testing our conjectures so as to arrive at causal laws or hypotheses describing what happens next in various initial conditions. Whether or not the world is governed by underlying forces and mechanisms, our knowledge of it, formulated in terms of causal laws, has only the empirical warrant conferred when observation and experiment uphold our generalizations. Scientific method is a matter of generalizing either from a known pattern to the next case (prediction) or from a particular case to a pattern which subsumes it (explanation). Prediction and explanation are thus two sides of the only epistemic coin, experience generalized. The economic world and economic knowledge are then construed on this natural science model.

This very basic empiricism needs supplementing in two ways, if it is to carry conviction. One is to give theory a more explicit role, which it plainly has in the practice of economics, without undermining the claims of prediction to be the only test of truth. A neat source is Friedman (1953), who associates economic theory with 'a language' and 'a body of substantive hypotheses'. The former is 'a set of tautologies' and 'its function is to act as a filing system'. The latter is 'designed to abstract essential features of a complex reality'. Whether the right features have been abstracted and included in the filing system depends solely on thee success of the resulting predictions. This echoes the Logical Positivists' distinction between analytic statements, whose truth relies on the meaning of terms and tells us nothing about the world, and synthetic statements, whose truth depends on the facts. Friedman's 'substantive hypotheses' serve to link pure theory (the 'filing system') to the world, while making sure that empirical facts wear the trousers.

The other supplement is Popper's (1963) account of science as conjectures and refutations. Reflecting that circular theories, like those of Freud and Marx, are always confirmed by experience in the eyes of their holders, he weakens the claim of empirical confirmation as a test of truth. Instead, it is falsifiability which separates science from pseudo-science. Formally, if hypothesis H implies observation O, then O does not prove H, but *not-O* refutes H. So a genuinely scientific theory must state possible empirical conditions in which it would be refuted. A causal law is an empirical hypothesis of sufficient scope and generality, which has risked refutation. In Popper's own eyes, he has made a radical break with traditional empiricism by undermining the value of induction. However, epistemologically, claims to knowledge still face the same old test and the facts of observation are still trumps. (This paragraph refers to Popper's best-known, classic account and not to his more recent writings.)

Later philosophy of science has raised serious difficulties for this empiricist approach to judging a theory. The Quine–Duhem hypothesis has it that a scientific theory is a web, which includes observation sentences and which can only be understood as a whole. Quine (1961), evoking Duhem (1914), argues that traditional empiricism relies on two untenable dogmas. One is that our five senses supply us with 'unvarnished news' as an objective and independent test of hypotheses. The other is that meaningful statements divide cleanly into analytic and synthetic (as defined above). Both dogmas are to be rejected. Our beliefs form a 'seamless web', a 'field of force which touches experience only at the edges'. At the moment of testing they 'face the tribunal of experience together' and, in assessing the verdict, we always have a choice of what to accept, revise or reject (including our own observations). In place of a single hypothesis H implying an unvarnished observation O, we have a set H_1H_2 ... etc. linked to another theory-laden set O_1O_2 ... etc. What we do about it, when we find that we cannot keep both sets, may be governed by criteria like parsimony, elegance or fertility, whose epistemic warrant is open to question and is far removed from any which relies on there being brute facts.

Economics is full of illustrations of the Quine–Duhem hypothesis. The 'money supply' or 'the general price level' are not brute facts. There are a number of different definitions of the money supply and ways of aggregating prices, and the choice of one rather than another will reflect theoretical considerations. The economist works with descriptions of data which already have theoretical order built into them, and this seems inescapable. Likewise, the joint testing of hypotheses is notoriously recognized in economics to be complicated by the role of *ceteris paribus* conditions. Predictions are always issued subject to certain *ceteris paribus* clauses, and if the prediction is falsified, the economist is often able to claim that this is because the *ceteris paribus* conditions were violated, particularly when, as in consumer theory, there are unobservable variables like preference orderings or utility functions to reckon with. So we find that empirical tests in economics are often indecisive. Few neoclassical economists, for example, seem willing to forsake the homogeneity assumption (the absence of money illusion) in consumer theory despite its frequent 'falsification'. Indeed, it looks rather as if, despite the common gestures of respect for Popper, economics is full of those circular theories which are always confirmed by experience in the eyes of their holders. The Quine–Duhem hypothesis is descriptively plausible.

This conclusion would not surprise Kuhn (1970), who argues that even the natural sciences have not progressed in the way expected by an empiricist methodology. Instead, their history is best understood as a discontinuous series of paradigms. By 'paradigms' he usually seems to mean the definitive current practices of the dominant scientific community and sometimes a more loosely specified set of currently shared presuppositions or world views. Paradigms rise and fall for many reasons, but empirical testing is never decisive. Rather, paradigms simply acquire more and more anomalies as contrary empirical evidence accumulates, and it is only when a new paradigm surfaces, which can incorporate the anomalies, that the anomalies become regarded as counter-examples. A paradigm shift then occurs. Feyerabend (1975) has continued this

process of dismantling empiricism, until it is unclear whether theory choice can be a rational process at all. This can be read as an invitation to a sociology of knowledge approach. If there is no good intellectual reason behind theory selection, then we must study the social pressures on and within the scientific communities, which influence the evolution of theory. In economics one might cite the suspicion in some quarters that the dominance of neoclassical economics owes much to its apparent support for a free enterprise system and that its mathematical sophistication is best understood as an exclusionary device typical of a closed profession. But a step into the sociology of knowledge is not the only option.

Lakatos (1978) tries to reaffirm the core of Popper by accepting that the units at stake are whole research programmes rather than single hypotheses, and that the core of the research programme is often defended against falsification by suitable adjustments to the auxiliary hypotheses or 'protective belt'. A research programme is to be judged, however, by whether these revisions to the protective belt are progressive or degenerating. Degenerating ones are *ad hoc* and cover only the anomaly which has precipitated the adjustment, whereas progressive ones provide additional and novel areas of application for the theory. The progressive/degenerative distinction sounds a promising way of reintroducing empirical criteria into the evaluation of a theory, but in practice this designation, like falsification earlier, is prone to vary with the eyes of the beholder. What looks progressive from one theoretical perspective can become *ad hoc* when viewed from another. For a setting to these developments the reader might usefully consult Harré (1972); for a chart of the options, Chalmers (1976); and for a robust post-empiricist philosophy of science, Hesse (1980).

Recent philosophy of science thus makes it unsurprising that economic methodologies inspired by empiricism are alive with controversy. But there are other sources of methodological inspiration. Weber (1922) draws an appealing but cloudy distinction between explaining *(erklären)* and understanding *(verstehen)*. Natural sciences seek to explain by means of causal laws; the humanities seek to understand by reconstructing the actors' world from within. The crux is the idea that the agent's own point of view matters in the social world in a way which does not hold for the natural sciences. Unlike the natural world, the social is *not* an ordered realm independent of our concepts, beliefs, hypotheses and conjectures about it. Our beliefs influence our actions and hence the outcomes we observe in any empirical investigation, whereas subatomic particles have no beliefs to affect outcomes in the natural world. This sets a puzzle for an empiricist methodology, which seeks the causal laws governing an independent realm. How is the social scientist to investigate the world from within and to relate these findings to the demands of objectivity?

Weber's answer is that some of his work is to be done by a process of *'verstehen'* – a key term from the German idealist, 'hermeneutic' (interpretative) tradition. The guiding thought is that what turns behaviour into action is its inward meaning and that institutions similarly are meaningful practices. But 'meaning' is an elusive concept, which threatens to let in more subjective variety than a social scientist can welcome. Weber tries to render *verstehen* more precise by stressing the rationality of action seen through the actor's eyes. He borrows the neoclassical economic concept of rational action. To the objection that real actions are not always rational, even when seen from within, he replies that, by establishing what would be fully rational, one can identify departures from the ideal type as *explananda*. *Verstehen* is not the only method, however. The social sciences seek both adequacy at the level of meaning and adequacy at the causal level, with the 'causal level' said to be one of statistically significant correlations. *Verstehen* is thus finally less of an alternative to *erklären* and more of a heuristic device; but an interesting line has been opened.

Among examples of the use of *verstehen* Weber cites pure mathematics. This suggests a more ambitious thought, one which finds echoes in economics itself. Von Mises (1949, 1960) presents economics as the science of human action and economic theory as the construction *a priori*

of ideal types of rational allocation. This yields an element of *a priori* knowledge contrary to empiricism. Similarly, Hayek (1960) and the new Austrians appeal to Kant and *a priori* knowledge when making their commitment to a methodological individualism grounded in the preconditions of the possibility of free choice.

Although a shift to a Kantian epistemology of pure reason would create more problems than it answers, it could have interesting implications. Consider, for example, Arrow–Debreu general equilibrium theory. It is not plausibly regarded either as a set of empirical hypotheses or as a mere filing system. Its proponents would not accept a sociology-of- knowledge suggestion that it belongs to the initiation rites of the profession. It functions very much as an 'ideal type', used for judging economic performance and making policy recommendations. What claim has to be made for it, if it is to be a reliable benchmark for performance? A Kantian reply is that it has to lay claim to truth, in the sense of stating correctly what would be the outcome of a fully rational allocation of resources throughout an economy. As in mathematics, the Kantian adds, *a priori* truths are hard to come by but are nonetheless what theory aims for. Arrow–Debreu general equilibrium theory makes sense on no other terms.

This is, of course, a contentious approach even to pure mathematics. It is doubly so for economics, owing to the role of rationality in its 'ideal types'. Not everyone would accept that Arrow–Debreu general equilibrium theory embodies an ideally rational allocation. Whether it does depends in part on how 'rationality' is defined. It is standardly given an instrumental definition, with a rational allocation being one which adopts the most efficient means to a given end. One reason for this definition is that it is believed to keep on the safe side of the positive/normative distinction and so to preserve the basic parts of economic theory from value judgements. This does not satisfy advocates of other approaches, especially political economists, who allege that Arrow–Debreu general equilibrium theorem has ideological commitments. A Kantian comment would be that 'ideal type' rationality is bound to involve the rationality of ends as well as of means. An ideally rational allocation of resources is the one which a just or good society would display, and an ideally rational choice is, in the last analysis, a moral choice. A switch to a Kantian epistemology may breach the positive/normative distinction and restore economics to the position of a moral science.

A challenge to the positive/normative distinction opens up deep epistemological questions. But some economists are willing to take the plunge, inspired by Rawls (1971). By defining a just society as one whose allocation of resources rational agents would agree upon in advance of knowing what they would each get out of it personally, Rawls connects economic rationality to moral choice. That makes it possible to ask precise questions about, for instance, the rationality of preferences and which kind of preferences should be satisfied in cases of conflict. The hope might be said to be an *a priori*, normative science with implications both for efficiency and for moral advance. This is to take seriously the thought that to know what would be ideally rational is to know how to do better.

Rawls's epistemological novelty lies in the use of a thought-experiment, inviting readers to think themselves into the shoes of fully rational, self-interested agents, who do not yet know whether they themselves will gain or lose from any arrangement proposed. Their knowledge of what would be rational comes from a proof that each does best to settle for equal basic rights and a maximum distribution of goods. The proof is controversial, but Rawls has certainly given economists new thoughts, especially in welfare economics, and a new line of defence against the charge that pure theorizing about reflective equilibria is only a parlour game.

The article began by suggesting that the lack of many results in economics comparable to the major discoveries in natural science reflected puzzles about causal laws and the value of seeking them. Some of the puzzles are general for all sciences; witness the unfinished arguments started by Quine, Kuhn and others. An initial empiricism seems peculiarly difficult to uphold in economics, however, because one is trying to predict what will be done by agents, who themselves have

beliefs and whose world depends on their expectations.

That makes rationality an epistemologically important yet special concept in economics. But it is unlikely to be a gateway to a simple rationalist epistemology. Whereas an 'ideal type' of frictionless motion is simply a limiting case with a zero coefficient of friction, an 'ideal type' of rational choice is not an abstraction from normal behaviour but a solution to a theoretical problem with a likely normative dimension. Even if economics is still regarded as the search for a different kind of causal law, rather than as an alternative to causal thinking, this difference in kind is great enough to set epistemological problems. Changes in belief change the course of economic events and introduce discontinuities, which make ideal types of rational action unlike timeless models of causal regularities. Rational belief is part of the concept of rationality. So the normative element of the concept of rationality will remain important for economics, even if only for prediction.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Economic Man
- ▶ Models and Theory
- ▶ Philosophy and Economics
- ▶ Rhetoric of Economics

Bibliography

Chalmers, A.F. 1976. *What is this thing called science?* St Lucia, Queensland: University of Queensland Press.

Duhem, P. 1914. *The aim and structure of physical theory*, 1954. Princeton: Princeton University Press.

Feyerabend, P.K. 1975. *Against method: An outline of an anarchistic theory of knowledge*. London: New Left Books.

Friedman, M. 1953. On the methodology of positive economics. In M. Friedman, *Essays in positive economics*, Chicago: University of Chicago Press.

Harré, R. 1972. *The philosophies of science*. Oxford: Oxford University Press. Hayek.

Hayek, F.A. 1960. *The constitution of liberty*. Chicago: University of Chicago Press.

Hesse, M. 1980. *Revolutions and reconstructions in the philosophy of science*. Brighton: Harvester Press.

Kuhn, T.S. 1970. *The structure of scientific revolutions*, 2nd edn. Chicago: University of Chicago Press.

Lakatos, I. 1978. The methodology of scientific research programmes. In *Philosophical papers*, Vol. I, ed. J. Worral and G. Currie, Cambridge: Cambridge University Press.

Lipsey, R.G. 1980. *Introduction to positive economics*. London: Weidenfeld & Nicolson.

Popper, K.R. 1963. *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge & Kegan Paul.

Quine, W. 1961. Two dogmas of empiricism. In *From a logical point of view*, ed. W. Quine. Cambridge, MA: Harvard University Press.

Rawls, J. 1971. *A theory of justice*. Cambridge, Mass.: Belknap Press of Harvard University Press.

von Mises, L. 1949. *Human action*. London: William Hodge.

———. 1960. *Epistemological problems of economics*. Princeton: Princeton University Press.

Weber, M. 1921. *Economy and society*, 1968. New York: Bedminster Press.

Equal Rates of Profit

Christopher Bliss

The concept of equality, or its opposite inequality, implies a comparison, and a comparison must be based on the consideration of a population of cases. Therefore equality or inequality has different implications according to the definition of that population. This general observation applies in particular to rates of profit.

Three different types of comparison of rates of profit will be examined:

(i) We may compare the rates of profit in terms of a fixed *numéraire*, particularly money, which can be obtained over a certain period of time from investment of funds in different lines of activity. We shall refer to equality in this sense as *sectoral equality* of rates of profit. Or,

(ii) We may compare the rate of profit obtainable over a certain period of time in terms of one *numéraire* with that obtainable over the same period of time in terms of another *numéraire*.

In a famous chapter of his *General Theory* (Keynes 1936, Ch. 17), Keynes employs the term 'own rates of interest' to describe these rates of return in terms of different *numéraires*, and we shall borrow the same term and call equality of the rates of return in different numeraires *own rates* equality. Finally,

(iii) We may compare the rate of profit obtainable in terms of the fixed *numéraire*, which may again be money, during one period of time with that obtainable during another later period of time. This comparison include the historically important question of the long-term trend of the rate of interest, whether it will tend to constancy, to increase, or to decline and, if not constant, what will be its eventual limit. We shall refer to equality in this sense as *temporal equality* of rates of profit. (In common with many writers, particularly in the past, we ignore in the present discussion distinctions between the rate of interest and the rate of profit. The main cause of a persistent difference between the two must be sought in the uncertainty from which our analysis abstracts.)

While it is convenient to have discussions of respectively sectoral, own-rate and temporal equality of rates of profit collected together in one article, it will be clear that these are distinct notions and that the investigation of the conditions required for another.

The Theory of Profit

An argument concerning equality of rates of profit might depend importantly on which theory of the rate of profit is invoked. Such is inescapably the case where temporal equality of rates of profit is concerned. However a good deal of our argument concerning sectoral and own rate equality of rates of profit is independent of the exact theory of the determination of rates of profit in general. This unexpected possibility might be realized because equality of rates of profit depends above all upon arbitrage, the tendency for capital to seek the highest return. Indeed in some cases an arbitrage condition alone suffices to demonstrate that rates of profit must be equal.

We shall refer to a state of the economy in which all possibilities of profitable arbitrage have been put into effects, which is a kind of short-period equilibrium, as an *arbitrage equilibrium*. It has sometimes been claimed that profit (where what is intended is a part of profit distinct from a normal rate of return) is essentially a phenomenon of disequilibrium. On this account an arbitrage equilibrium would not only exhibit equal rates of profit, all rates of profit would equal zero. Only the normal rate of return would be realised in an arbitrage equilibrium.

To argue about terminology where weighty issues are involved shows poor judgement. Even if profit is defined to be an excess of return to capital above the return generally available, and even if we exclude temporary rents, it remains to show that no sector can enjoy a permanent profit advantage against which arbitrage is for some reason powerless. If, on the other hand, profit is taken to include temporary rents it is evident that there is really no case for equality. Hence the only interesting question to decide is whether rates of profit defined as net returns to capital divided by the values of capital employed (on average or at the margin) are equal in an arbitrage equilibrium.

Sectoral Equality of Rates of Profit

Nowhere is the power of arbitrage, together with its limitations, better illustrated than in the case of comparisons of profit rates across sectors. The desire of every investor to obtain the highest possible rate of return may reasonably be assumed to equalize the equivalent rates of return on different bonds. Will not a similar principle ensure the equalization of rates of profit in different activities, be they regions or industries?

The answer depends on two important points. First, we must decide how to compare two rates of return, what are the principles of equivalence? Secondly, arbitrage may encounter obstacles. This is true even where bonds are concerned,

and is more important still where we are concerned with different sectors.

Clearly rates of return should be true economic rates including allowances for capital gains, etc. Moreover, two apparently different rates of return may not excite arbitrage if they represent different risks, or different liabilities to taxation, or if the difference is too small to overcome transactions costs. Although they are important in empirical investigations, these detailed considerations may be neglected for our purposes. So we are left with structural obstacles to arbitrage.

When economic theorists assume equal rates of profit in different sectors they are implicitly ignoring questions of industrial structure. (For an excellent treatment of the concept of industrial structure and it implications for profitability, see Hay and Morris 1979, Ch. 7.) It is typically supposed, for example, that capital may be shifted from one sector to another in arbitrarily small quantities. If increasing returns to scale imply that operation at a very small scale will be costly, the putative entrant must choose between staying out of the sector of fighting his way into what must be an oligopolistic market. There is naturally no reason to suppose that the rate of profit enjoyed by those already inside may not exceed that obtainable in a competitive sector of small-scale units.

It would not be necessary to reiterate the foregoing point if it had not apparently been challenged by the late Piero Sraffa in the oft-quoted foreword to his *Production of Commodities by Means of Commodities* (1960). Sraffa's model for the determination of prices is striking for its simplicity and for the fewness of its assumptions. In his forward the author warned his readers against the temptation to assume that his argument depended upon assuming constant returns to scale. In a sense it does not, as that assumption is never directly employed. However equality of rates of profit, sectoral equality according to our present terminology, is assumed. We cannot of course claim that sectoral equality requires constant returns to scale. However it requires some assumptions about the environment, specifically the market environment, in which firms operate, and constant returns to scale and free entry are obvious sufficient conditions for sectoral equality of profit rates.

Equality of Own Rates of Interest

Consider a price system extending through time so that for each period t there is a present price for each of N goods. Such a price system may be represented thus:

$$\begin{matrix} p_{11} & p_{12} & \cdots & p_{1t} & \cdots & p_{1T} \\ p_{21} & p_{22} & \cdots & p_{2t} & \cdots & p_{2T} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{N1} & p_{N2} & \cdots & p_{Nt} & \cdots & p_{NT} \end{matrix} \quad (1)$$

As problems raised by infinite price systems need not concern us here, we suppose that the prices only extend forward to period T. If we imagine that good 1 is money, it will be seen that the money rate of interest in period 1 for a t-period loan may be calculated as follows. One unit of present money costs p_{11} and one unit of money bought now for delivery in period t costs p_{1t}. Hence one unit of money surrendered now buys p_{11}/p_{1t} units of money at t. This corresponds to a rate of interest equal to $p_{11}/p_{1t} - 1$, or $(p_{11} - p_{1t})/p_{1t}$. What was denoted above by the term the money rate of interest can equally be designated the own rate of interest on money, in this case for a t-period loan.

The money rate of interest measures the extra money obtainable by postponing payment as a proportion of the payment deferred. This notion generalises to any good. We may for example measure the extra wheat obtainable by postponing delivery as a proportion of the quantity of wheat delivery deferred. Suppose that wheat prices occupy the second row of (Eq. 1) above. Then the t period own rate of interest for wheat will be equal to $p_{21}/p_{2t} - 1$, or $(p_{21} - p_{2t})/p_{2t}$, which is exactly analogous to the expression of the money rate of interest already derived.

Turning from the rows of (Eq. 1), which correspond to different goods, consider the columns, which correspond to different periods of time. It is easily shown that if the columns are proportional to each other, which is the same as saying that relative prices are the same in all periods, then the own rate of interest for a given duration of loan is the same for all goods. Suppose that the own rate

of interest for good 1 for a deferment from period 1 to period t is r_1t. Then, as we have already seen:

$$r_1t = (P_{11} - p_{1t})/p_{1t} \quad (2)$$

However, by assumption:

$$p_{1t}/p_{1t} = p_{11}/p_{1t} \quad (3)$$

So that:

$$(p_{11} - p_{1t})/p_{1t} = (p_{11} - p_{1t})/p_{1t} \quad (4)$$

Or,

$$r_1t = r_1t \quad (5)$$

Here, constancy of relative prices implies equality of own rates of return, as required. Conversely, variations in relative prices will be reflected in differences in own rates of return.

Under what circumstances is it reasonable to assume constancy of relative prices over time? We shall certainly require the assumption that the economy is stationary in some sense. Suppose for example that as time passes timber becomes more and more scarce relative to demand as forests are depleted or demand grows. Then we would expect the price of timber to rise through time relative to other goods. Similarly, technical progress, unless it be of the simplest labour-augmenting kind, will typically imply changes in relative prices. The transistor, the microchip and other innovations, to cite another example, have certainly caused electric goods to become relatively cheaper.

Consider therefore a stationary state, which may be growing economy, but which is stationary in the sense that in each period it is technically exactly the same as in every other period, except perhaps for scale. As the economy is essentially the same at every moment of time, it makes intuitive sense to suppose that relative prices might be the same at each moment of time, and this intuition is valid in so far as it can be shown that any development of the economy which is stationary, in the sense just described, may be supported by a price system which is itself stationary, in the sense that relative prices are invariant over time[3].

Stationarity of the real economy is sufficient for stationarity of a price system that will support production activities, but does not imply that any such price system will be stationary. Indeed it is an implication of the multiplicity of price systems and interest rates which goes under the name of 'double-switching'. That prices which support stationary production will frequently be neither unique nor themselves stationary. Their non-uniqueness is an immediate implication of double-switching. The existence of non-stationary price systems for these equilibria follows when we note that the average of two systems of equilibrium prices must themselves be equilibrium prices. However the average of two price systems based on different rates of interest produces a rate of interest variable over time, and varying relative prices.

The importance of these findings may be questioned because the price system is required to support not only production (supply) but also consumption (demand). This will make the observation of non-unique prices, and in particular of a history including double-switching, much less probable than a considering of the production side alone might suggest.

It remains to briefly mention Keynes's use of own rates of interest in his *General Theory*, if only to point out that it is not in fact particularly germane to the present discussion of equality of own rates of interest. Keynes's extraordinary argument is concerned with the comparison of money rates of return at the margin to accumulating various assets, which is something like the question of sectoral equality.

We may imagine that as the various assets are accumulated the money rates of return to further accumulation for each of them is forced down, and that the quantities accumulated are such that these marginal returns on all assets are equalized. If we could conceive of the elasticity of the money rate of return for each asset to the stock accumulated (which we may call the return-stock elasticity) as a value independent of other accumulations, which Keynes in effect does, then assets with low return-stock elasticities will accumulate rapidly relatively to assets with higher return-stock elasticities. Keynes's argument claims that money is eventually the asset with the lowest return-stock elasticity, and that this has the implication that, in an economy with a limited supply of money, the money rate of return

(which of course is the own rate of interest of money) will eventually rise to a level which discourages the further accumulation of real assets.

Temporal Equality of Rates of Proit

We now turn to the equality, or inequality as the case may be, of the rates of profit which prevail at different moments of time. There is a longer tradition among economic theorists, which goes back to the classical writers, of explaining the long-run tendency for the rate of profit to fall. This was largely a response to a supposed fall in the rate of interest which the classical economists 'took to be an indisputable fact'. For these older theories the reader is referred to entries on Adam Smith, Marx, Mill, Ricardo and Say. Here we consider only a modern view of the problem. A justification for this division of labour may be sought in the fact that modern theories of the rate of profit are radically different from classical views.

The main source of the difference between modern and classical theories (which in this context should be taken to exclude Marx) is that the former treat technical progress as having regular and continuous effects on the economy, where the latter typically do not. Thus the characteristic classical argument for a falling rate of profit is stagnationist in nature. The decline in the rate of profit is part of the grinding to a halt of a previously progressive economy. In contrast, the modern neoclassical approach locates the explanation of a falling rate of profit in the character of a technical progress conceived as an indefinitely continuing process.

To demonstrate the theoretical issues involved we first show when a declining rate of profit would arise in a neoclassical model with aggregate capital and a constant saving propensity, and then discuss some of the shortcomings of that model as an account of capital accumulation.

Let output, Y, depend upon the input of labour, L, and a capital stock which is homogeneous with the output flow, K, according to a constant returns production function as:

$$Y = F(K, L, t). \tag{6}$$

Let partial derivatives be denoted by subscripts so that, for example, the marginal product of capital is denoted $F_K(K, L, t)$. We denote the rate of profit by r, so that:

$$r = F_K(K, L, t). \tag{7}$$

Time derivatives are shown by a dot over the variable concerned. Differentiating $F_K(K, L, t)$ totally with respect to time we obtain an expression for the time rate of change of the rate of profit as:

$$\dot{r} = F_{KK} \cdot \dot{K} + F_{KL} \cdot \dot{L} + F_{Kt}, \tag{8}$$

Hence for constancy of the rate of profit we must have:

$$F_{KK} \cdot \dot{K} + F_{KL} \cdot \dot{L} + F_{Kt} = 0. \tag{9}$$

which on rearrangement yields:

$$\frac{F_{KK} \cdot K}{F_K} \cdot k + \frac{F_{KL} \cdot L}{F_K} \cdot l + \frac{F_{Kt}}{F_K} = 0, \tag{10}$$

where k and l are respectively the logarithmic rates of growth of capital and labour. Now (Eq. 5) can be expressed more simply as:

$$\sigma_K \cdot k + \sigma_L \cdot l + \gamma = 0; \tag{11}$$

where σ_K and σ_L are respectively the elasticity of the marginal product of capital with respect to K and L, and y is the proportional change in the marginal product of capital due to the passage of time alone. We know that $F_K(K, L, t)$ is homogeneous of degree zero in K and L. Hence:

$$\sigma_K + \sigma_L = 0, \tag{12}$$

and (Eq. 11) reduces to:

$$\sigma_K \cdot (k - 1) + \gamma = 0. \tag{13}$$

This last expression has an intuitive interpretation. As σ_k is the elasticity of the marginal product of capital with respect to capital, it will be negative. It is weighted by $k - 1$, the rate of growth of capital per unit of labour, which will be positive under

normal economic growth. Thus $\sigma K \cdot (k - 1)$ measures the rate at which capital accumulation is pushing down the rate of profit due to the substitution of capital for labour at constant technical knowledge. The second term represents the rate at which technical progress is tending to raise the rate of profit at constant factor proportions, which must be positive term if technical progress is beneficial. Now, unsurprisingly, (Eq. 13) says that, for the rate of profit to remain constant, these two effects must exactly offset.

As it is known that a production function with aggregate capital cannot be derived rigorously except for simple or special production technologies, it may reasonably be asked how fare the above account, of a downward pressure on the rate of profit due to accumulation being offset by an upward pressure due to technical progress, generalizes. In particular, is it generally true that accumulation with constant technical knowledge exerts a downward pressure on the rate of profit?

Given the enormous literature on the theory of capital which has been produced in recent years, it is perhaps surprising that this question remains relatively under investigated. Many discussions of capital accumulation simply beg the question by assuming that the rate of interest would fall continuously through time. Indeed double-switching is most at variance with the traditional neoclassical view of capital accumulation when that assumption is made. However there is no guarantee of a continuous fall of the rate of profit through time, and the demand side of the economy is likely to prohibit a return to a previous and lower income state.

On the other hand, linear models of the type that have been used to illustrate simple stories of capital accumulation can lead to quite eccentric time profiles of consumption being associated with the accumulation of capital (where this is defined simply as an increase in long-term consumption). Hence there is no possibility in general of ruling out erratic developments in the rate of interest over time.

See Also

- ▶ Capital Perversity
- ▶ Development Economics
- ▶ Interest and Profit
- ▶ Surplus Approach to Value and Distribution
- ▶ Sraffian Economics

Bibliography

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam: North-Holland.

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

Hay, D.A., and D.J. Morris. 1979. *Industrial economics: Theory and evidence*. Oxford: Oxford University Press.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

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

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

E

Equality

James S. Coleman

Abstract

'Equality' is used to mean equality before the law, equality of opportunity, and equality of result, among other things. These types of equality are not necessarily mutually compatible. Equal distribution of benefits is often taken to be 'natural' (by Rawls, for example), partly because envy is ubiquitous. In welfare economics the presumed diminishing marginal utility of money implies that equality of incomes maximizes welfare, but if interpersonal utility comparisons are impossible no such presumption can be made. As well, the interdependencies between individuals in terms of welfare are such that enforced equalization is likely to reduce overall welfare.

Keywords

Berlin, I.; Coleman, J. S.; Edgeworth, F. Y.; Efficiency vs. equality; Envy; Equality; Equality before the law; Equality of opportunity; Equality of result; Inequality; Input–output

analysis; Interdependencies of welfare; Interpersonal utility comparisons; Liberty; Marginal utility of money; Nozick, R.; Optimal taxation; Pigou, A. C.; Progressive taxation; Property rights; Rawls, J.; Robbins, L. C.; Welfare economics

JEL Classifications
D6

The very use of the term 'equality' is often clouded by imprecise and inconsistent meanings. For example, 'equality' is used to mean equality before the law (equality of treatment by authorities), equality of opportunity (equality of chances in the economic system), and equality of result (equal distribution of goods), among other things. These different meanings often conflict, and are almost never wholly consistent. See Hayek (1960, p. 85, 1976, pp. 62–4) for a discussion of equality before the law and equality of result, and Rawls (1971) for a discussion of equality of opportunity within a theory of distributive justice. Elsewhere I have discussed the difference between equality of opportunity and equality of result in education (Coleman 1975). See also Pole (1978) for a detailed examination of the changing conceptions of equality in American history.

Some order can be brought into the confusion among the different uses of the term 'equality' by first conceiving of a system that constitutes an abstraction from reality. The system consists of:

(a) a set of positions which have two properties:

(i) when occupied by persons, they generate activities which produce valued goods and services;
(ii) the persons in them are rewarded for these activities, both materially and symbolically;

(b) a set of adult persons who are occupants of positions;
(c) children of these adults;
(d) a set of normative or legal constraints on certain actions.

What is ordinarily meant by equality under the law has to do with (b), (c), and (d): that the normative or legal constraints on actions depend only on the nature of the action, and not on the identity of the actor. That is, the law treats persons in similar positions similarly, and does not discriminate among them according to characteristics irrelevant to the action.

What is ordinarily meant by equality of opportunity has to do with (a), (b), and (c): that the processes through which persons come to occupy positions give an equal chance to all. More particularly, this ordinarily means that a child's opportunities to occupy one of the positions (a) do not depend on which particular adults from set (b) are that child's parents. What is ordinarily meant by equality of result has to do with (a.ii): that the rewards given to the position occupied by each person are the same, independent of the activity.

These three conceptions, equality under the law, equality of opportunity, and equality of result can also be seen as involving different relations of the State to the inequalities that exist or spontaneously arise in ongoing social activities. Equality before the law implies that the laws of the State do not recognize distinctions among persons that are irrelevant to the activities of the positions they occupy, but otherwise make no attempt to eliminate inequalities that arise. Equality of opportunity implies that the State intervenes to insure that inequalities in one generation do not cross generations, that children have opportunities unaffected by inequalities among their parents. Equality of result implies a continuous or periodic intervention and redistribution by the State to insure that the inequalities which arise through day-today activities are not accumulated, but are continuously or periodically eliminated.

The relations between the first two kinds of equality differ according to how close a society is to a legally minimalist society or a legally maximalist society. In a society that is legally minimalist, equality before the law is compatible with a high degree of inequality of opportunity – depending on the distribution of opportunity provided by other institutions in society, such as the family. In a legally maximalist society, in which many functions of traditional institutions have

been taken over by institutions that are creatures of the State (e.g., functions of the family taken over by the public school), equality before the law implies a high degree of equality of opportunity. Only in a society in which the law was far more intrusive than found anywhere, and children were taken from their families to be raised 'with equal opportunity' by the State, could it be said that equality before the law would coincide with equality of opportunity.

The relation between equality of opportunity and equality of result is somewhat different, for it implies two different kinds of interventions of the State. Equality of opportunity implies intervention to provide each person with resources that give equal *chances* to obtain the material and symbolic rewards that arise from productive activity, while equality of result implies intervention in the distribution of these rewards, to provide each person with equal *amounts.* The two concepts become indistinguishable only when the State intervenes to insure that each position (in (*a*) above) provides the same set of material and symbolic rewards; and in such a circumstance, 'opportunity' loses meaning altogether.

Is Equality 'Natural'?

There are certain philosophical positions that take equality of result as a 'natural' point, from which all others are deviations. Isaiah Berlin probably states this as well as any other

> No reason need be given for . . . an equal distribution of benefits for that is 'natural', self evidently right and just, and needs no justification, since it is in some sense conceived as being self justified . . . The assumption is that equality needs no reasons, only inequality does so; that uniformity, regularity, similarity, symmetry, . . . need not be specially accounted for, whereas differences, unsystematic behavior, changes in conduct, need explanation and, as a rule, justification. If I have a cake and there are ten persons among whom I wish to divide it, then if I give exactly one tenth to each, this will not, at any rate automatically, call for justification; whereas if I depart from this principle of equal division I am expected to produce a special reason. It is some sense of this, however latent, that makes equality an idea which has never seemed intrinsically eccentric . . . (1961, p. 131).

This quotation describes a view with which Berlin does not necessarily identify himself. In the same paper, he states that 'equality is one value among many ... it is neither more nor less rational than any other ultimate principle ... rational or nonrational' It is, however, the position implicitly taken by John Rawls in his *Theory of Justice,* for the book is addressed to the question, 'When can inequalities (of result) be regarded as just?' Rawls's answer can be paraphrased as 'Only those inequalities are just which make the least well off person better off than that person would be (other things being equal) in the absence of the inequalities.'

Whether equality of result is 'natural' or not, and whether the position of Berlin and Rawls is correct or incorrect, would appear to depend on how the distribution of goods occurs: If goods are initially the property of a single central source (e.g., 'the State'), then Berlin's position and that of Rawls appear correct. If all rights and resources originate with the State (or with the king, as in early political theory), than an equal distribution has some claim to be seen as natural. (If, for example, the revenue from oil discovered on public lands is a major component of GNP, as in some Middle Eastern states, equal distribution constitutes a natural point.) But if goods are seen to arise from the activities of a set of independent actors each with certain initial property rights, and each with a certain amount of zeal and skill, 'equality' (meaning equality of result) is hardly natural, and is inconsistent with the distribution of property rights including rights to the fruits of one's own activity.

Equality, Envy and Resentment

The idea of equality as 'natural' appears also to derive in part from the ubiquity of envy and resentment in society, with the demand for 'equality' as an expression of these feelings which carries legitimacy. A number of sociologists have pointed to this connection. For example, Simmel writes (1922, translated in Schoeck 1969, pp. 236–7):

> Characteristically, no one is satisfied with his position in relation to his fellow beings, but everyone

> wishes to achieve a position that is in some way an improvement. When the needy majority experiences the desire for a higher standard of living, the most immediate expression of this will be a demand for equality in wealth and status with the upper ten thousand.

Simmel follows with an anecdote: at the time of the 1848 revolution, a woman coal-carrier remarked to a richly dressed lady, 'Yes, madam, everything's going to be equal now; *I* shall go in silks and you'll carry coal.'

Helmut Schoeck, in an extensive examination of the role of envy in society, argues that

> social philosophers have largely failed to see how little the individual is concerned with being *equal* to someone else. For very often his sense of justice is outraged by the very fact that he is denied the measure of inequality which he considers to be right and proper (1969, p. 234).

Feelings of envy and resentment constitute a challenge to the existing distribution of *rights* in society, between those held collectively and those held individually. In particular, it is a challenge to the existence of individual property rights. The centrality of property rights for conceptions of equality is seen most clearly in neoclassical economic theory, which assumes a distribution of property rights among a set of independent actors, accompanied by a free market. (See Meade 1964, for a discussion of property rights and the market in relation to equality.) It is to economic theory that I now turn.

The Role of 'Equality' in Economic Theory

The concept of 'equality' has no place in positive economic theory. In this it is unlike the concept of 'liberty', for economic theory is predicated on the assumption of liberty, that is, free choice (subject only to resource constraints) among alternative actions. There is, in the concept of free choice, however, something closer to the idea of equality before the law than to equality of opportunity, and closer to the latter than to equality of result. Equality of result implies a distribution process that is the antithesis of the market.

But normative economics, that is, welfare economics, makes up for the absence of 'equality' from positive economic theory, for the idea of equality of result is a part of the very atmosphere surrounding welfare economics. The question of what policies will maximize social welfare is not often answered directly in terms of equality in the distribution of valued goods, but the idea seems always to hover nearby. The most direct expression of the central importance of equality in welfare economics was probably that of Pigou (1938); see also (Bergson (1966, ch. 9) who reasoned that because money, like everything else, had declining marginal utility, and thus a dollar was worth much less to a person when he had a million others than when it was the only one he had, then the maximum of social welfare could only be achieved when incomes were made equal. (Neither Pigou nor any other welfare economist followed this implication with actual policy recommendations for equality of income, thus raising the question: if the criterion is correct, then why not recommend implementing it?)

The rock on which Pigou's argument is often regarded as foundering is that of interpersonal comparison of utility. To move from the relative importance for one person of a dollar when he is rich and when he is poor to its relative importance to different persons is a move which, as has been often reiterated, cannot be justified on positive grounds. Perhaps the most widely quoted statement to this effect is that of Lionel Robbins (1938):

> But, as time went on, things occurred which began to shake my belief in the existence between so complete a continuity between politics and economic analysis . . . I am not clear how these doubts first suggested themselves; but I well remember how they were brought to a head by my reading somewhere – I think in the work of Sir Henry Maine – the story of how an Indian official had attempted to explain to a high-caste Brahmin the sanctions of the Benthamite system. 'But that,' said the Brahmin, 'cannot possibly be right – I am ten times as capable of happiness as that untouchable over there.' I had no sympathy with the Brahmin. But I could not escape the conviction that, if I chose to regard men as equally capable of satisfaction and he to regard them as differing according to a hierarchial schedule, the difference between us was not one which could be resolved by the same

> methods of demonstration as were available in other fields of social judgement … 'I see no means,' Jevons had said, 'whereby such comparison can be accomplished.'

Edgeworth expressed the same point, 'The Benthamite argument that equality of means tends to maximum happiness, presupposes a certain equality of natures; but if the capacity for happiness of different classes is different, the argument leads not to equal, but to unequal distribution' (1897, p. 114).

Such arguments are ordinarily taken as conclusive within the domain of economics, and with their acceptance, the very programme of welfare economics – not to speak of the foundations for a policy designed to bring equality – is emasculated.

A philosopher might argue, of course, that there is no logical difference between the comparison of utilities of two persons and the comparison of utilities of one person at two different times. Neither, by this argument, is warranted. See, for example, Parfit (1984).

However, Pigou's conclusion has, quite apart from problems of interpersonal comparison of utility, another deficiency. It assumes that each person is an island, and contributes nothing to the welfare of others, nor has his welfare contributed to by others. Yet is is the essence of social and economic systems that there is interdependence, that one person's activities do affect the welfare of others, whether intended or not. One person spends money on loud radios that cause disturbance, while another plants flowers that others enjoy. Or one uses income for training which is productive, benefiting general welfare, while another uses income on drink and becomes alcoholic, requiring public-expense hospitalization.

But if this is so, then maximization of welfare one time period into the future would require that these interdependencies be taken into account. Maximization would occur only if resources were distributed among persons in accordance with the positive impact of their activities on those events which bring welfare to others. But in general persons do not capture the full benefits of their welfare-generating activities, nor do persons pay the full costs of their welfare–diminishing activities.

The matter can also be seen as a problem in input–output economics: What current allocation of resources among productive activities (i.e., among positions in the system as described earlier) will achieve some desired distribution of final consumption? If the aim is to maximize the sum of final consumption ('maximizing welfare'?), it is quite unlikely that either the current allocation necessary to achieve that, or the distribution of final consumption itself, will approach equality. Even if the desired final distribution is equality, and even if that is achievable within the system of activities, it is highly unlikely that the allocation at time *0* necessary to achieve that at time *t* will be equal. And it may well be that the only distribution at time *0* that would achieve equality at time *t* would do so at a low level of welfare, with each having less than if there were inequality at time *t* resulting from a different distribution at time *0*. If Pareto optimality is taken as a self-evident necessary condition for optimal policies, then because of the processes described above, a criterion of equal distribution (either initially or subsequently) would violate the condition. This suggests that Rawls's question was misdirected, and should have been 'when (assuming non-violation of constitutional rights) is *equality* of distribution justified?' and should have been answered, 'Only when there is no unequal distribution that would subsequently make each better off.'

Thus even if Pigou's point that maximizing welfare requires equalizing marginal utilities is accepted, and noncomparability of utilities is ignored, the policy implication of equalizing incomes appears shortsighted in the extreme. Another way of seeing so is by use of Robert Nozick's Wilt Chamberlain example, an example designed to argue against theories of distributive justice which, like that of Rawls, use the resulting distribution of goods ('end state theories', to use Nozick's term) as a criterion.

> Now suppose that Wilt Chamberlain is greatly in demand by basketball teams, being a great gate attraction. (Also suppose contracts run only for a year, with players being free agents.) He signs the following sort of contract with a team: In each home game, twenty five cents from the price of each ticket of admission goes to him. (We ignore the question of whether he is 'gouging' the owners, letting them

> look out for themselves.) The season starts, and people cheerfully attend his team's games; they buy their tickets, each dropping a separate twenty five cents for their admission price into a special box with Chamberlain's name on it. They are excited about seeing him play; it is worth the total admission price to them. Let us suppose that in one season one million persons attend his home games, and Wilt Chamberlain winds up with $250,000, larger even than anyone else has. Is he entitled to this income? (Nozick 1974, p. 161)

Thus as Nozick points out, an equal distribution at one point will lead to an unequal distribution at a later point, due to the very system of activities through which persons satisfy their interests.

There are only three ways to prevent this, all of which, carried to their limit, can be shown to reduce welfare. One is to prevent the economic exchange through which persons spend their quarters as they see fit, for such exchanges may lead to a large accumulation in the hands of the Wilt Chamberlains.

A second is to attack the system of activities itself, the system which generates that matrix of coefficients that transform equality into inequality – that is, shutting down professional basketball, which redistributes income from those with low incomes to those with high incomes. The third way is to allow the exchange, but then to tax the high incomes back down to equality. This effectively eliminates the activity, because if income is an incentive to carry out the activity that is paid for, the Wilt Chamberlains lose all incentive to carry the activity.

Indeed, unless there is a perfect positive correspondence of those activities which are intrinsically pleasurable with those which produce benefits for others, and a perfect negative correspondence with those that produce harm for others, the absence of any extrinsic incentives will lower the welfare for all. The more interrelated the activities of individuals, the greater the reduction in social welfare when extrinsic incentives are absent.

It is true that taxation which is not carried to the limit, but is merely 'progressive', does not eliminate the incentive for activities that bring high income, for these activities continue in societies that have progressive taxation. But this taxation may lead to underprovision of welfare-generating activities. That is, efficiency may be sacrificed to achieve some distributional goals. The potential conflicts between efficiency and equality are discussed in the literature on optimal taxation (e.g., Atkinson and Stiglitz 1980, part II). (A device which is informally used in social systems to reduce the disincentive effect of regimes of taxation and redistribution that shift incomes in the direction of equality is the attachment of social stigma to the receiving of income thus redistributed, for example, stigma associated with being 'on welfare'. The existence of this stigma constitutes a means of informally reconstituting the differential incentives that are reduced by redistribution.)

All three approaches to preventing inequalities from arising out of equality give, at their extreme, the same result: elimination of the very system of activities that generates welfare in the first place; for it is these activities which not only generate welfare, but also transform equality at one time into inequality at a later time.

Thus it becomes clear that the source of inequalities is embedded in the very matrix of social and economic activities through which individuals increase the welfare of themselves and one another. If, through technology for example, this matrix changes in such a way that individuals' satisfaction of wants is more concentrated in a few hands (e.g., by the invention and development of television), then inequalities will necessarily increase.

More generally, the degree of inequality seems related to the degree of interdependence in this matrix of social and economic activities. In a social system that has very low interdependence (e.g., a social system composed largely of subsistence farmers, a condition that was once the case for nearly all societies), the welfare of each in future periods depends largely on his own initial distribution of resources (including zeal and skill). If that distribution is near equality, then near equality is perpetuated into the future, modified only by random events. More important, even if the initial distribution is unequal, the low interdependence of the system of activities

means that these inequalities (also modified by random events) are merely carried forward into the future. In a system with a high degree of interdependence, however, there are a great many configurations which constitute 'inequality-generating' activity structures. In such activity structures, initial distribution of equality will lead to highly unequal distributions. This inequality in turn will lead in the next generation to inequality of opportunity, constrained only by random processes or explicit policies towards non-inheritance of position, i.e., toward equality of opportunity. (In a system in which attention to basketball was directed not to televised professional teams, but to games of the local high school, both the material and nonmaterial rewards among basketball player would be more equally distributed. There would be greater equality of results, which would arise not through a change in the set of persons (*b*), the distribution of children (*c*), or the normative and legal constraints (*d*), but only through a change in the distribution of positions.)

Does this mean that there tends to be a negative relation between the interdependence of activities in a social system (and thus the total social product) and the equality with which the activities of the system distribute the product? If so, this is a discouraging result for those who would prefer a social system in which incomes are not increasingly unequal, for it specifies an opposition between two goals both regarded as desirable.

This question has two parts, a within-generation part and a between-generation part. Within generations, it appears likely that there is a negative relation, that increased interdependence does, except in unlikely activity structure, increase inequality. It is possible that this negative relation is responsible for the rise in redistributive actions of governments as interdependence of economic activities increases.

Between generations, the answer would appear to hinge largely upon the relative rates of increase of interdependence of activities and of equality of opportunity (i.e., non-inheritance of position). The latter can occur through regression to the mean as well as through explicit policy intervention (see Becker and Tomes 1986, for a discussion). If equality of opportunity increases more slowly than interdependence of activities, then (except for unlikely configurations of the activity matrix) there will be a decrease in equality of result among lineages of persons. If equality of opportunity increases more rapidly than the increase in interdependence of activities, there will be an increase in equality of result among lineages, even with a decrease in equality of result within generations.

Altogether, there has been little investigation of the matters discussed above, that is, just how the structure of social and economic activities itself affects inequalities. Such investigations would lead toward taking work on equality partly out of the realm of normative theory, bringing it partly into the realm of positive theory.

See Also

► Poverty

Bibliography

Atkinson, A.B.., and J.E. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.

Becker, G., and N. Tomes. 1986. Inequality, human capital, and the rise and fall of families. In *Approaches to social theory*, ed. S. Lindenberg, J. Coleman, and S. Nowak. New York: Russell Sage.

Bergson, A. 1966. *Essays in normative economics*. Cambridge, MA: Harvard University Press.

Berlin, I. 1961. Equality. In *Justice and social policy*, ed. F.A. Olafson. Englewood Cliffs: Prentice Hall.

Coleman, J. 1975. What is meant by 'an equal educational opportunity'? *Oxford Review of Education* 1 (1): 27–29.

Edgeworth, F.Y. 1897. *Papers relating to political economy*, 1925. London: Macmillan for the Royal Economic Society.

Hayek, F.A. 1960. *The constitution of liberty*. Chicago: University of Chicago Press.

Hayek, F.A. 1976. *Law, legislation and liberty*. Vol. 2. Cambridge: Cambridge University Press.

Meade, J.E. 1964. *Efficiency, equality and the ownership of property*. London: Allen & Unwin.

Nozick, R. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Parfit, D. 1984. *Reasons and persons*. Oxford: Oxford University Press.

Pigou, A.C. 1938. *Economics of welfare*. 4th ed. New York: Macmillan.

Pole, J.R. 1978. *The pursuit of equality in American history*. Cambridge: Cambridge University Press.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Robbins, L. 1938. Interpersonal comparisons of utility. *Economic Journal* 48: 635–641.
Schoeck, H. 1969. *Envy: A theory of social behavior*. New York: Harcourt, Brace and World.
Simmel, G. 1962. *Soziologie*. 2nd ed. Munich: Duncker & Humblot.

Equality of Opportunity

John E. Roemer

Abstract

Whereas the ethic of equality of outcome does not hold individuals responsible for actions that may create inequality of outcomes, equality of opportunity 'levels the playing field' so that all have potential to achieve equal outcomes; inequalities of outcome that then transpire are not compensable at the bar of justice. The influences on the outcome a person experiences comprise *circumstances* (for which he should not be held responsible) and *effort* (for which he should be). Equal-opportunity policy compensates persons for their disadvantaged circumstances, ensuring that, finally, only effort counts in achieving outcomes.

Keywords

Affirmative action; And parameters of disadvantage; Compensation; Distribution; And effort vs circumstances; Dworkin, R.; On equality; On insurance market; Educational finance; Efficiency; And equity; Equality of opportunity; Vs equality of outcome; Equality of outcome; Vs equality of opportunity; Vs equality of resources; And compensation; And efficiency; Income-tax regime; Insurance market; Dworkin on; Meritocracy; Property rights; And welfarism; Welfarism; Sen on; And property rights; Sen, A.; And welfarism

JEL Classifications

D63

Equality of opportunity is to be contrasted with equality of outcome. While advocacy of equality of outcome has been traditionally associated with left-wing political philosophy, equality of opportunity has been championed by conservative political philosophy. Equality of outcome does not hold individuals responsible for imprudent actions that may, absent redress, reduce the values of the outcomes they enjoy, or for wise actions that would raise the value of the outcomes above the levels of others'. Equality of opportunity, in contrast, 'levels the playing field' so that all have the potential to achieve the same outcomes; whether in the event they do depends upon individual choice.

This traditional political alignment was upset by Ronald Dworkin (1981a, b) who posed the question: if one is egalitarian, then what should one seek to equalize, welfare or resources? He argued, first, that equalizing welfare (outcome) was undesirable, even if interpersonal comparisons of welfare could be made, because doing so would fail to hold individuals accountable for their preferences. The issue of 'expensive tastes' was important for Dworkin; he argued that, if a person were glad he possessed an expensive taste, or identified with it, as opposed to viewing it as an addiction – a taste he would prefer not to have – then society owed him no extra resources to satisfy it. Dworkin argued that egalitarians should advocate the equalization of resources, as opposed to outcomes, but his conception of what comprised resources was broad. Resources consisted in not only transferable goods and wealth but internal talents as well. The question became: what allocation of *transferable* resources would count as equalizing the *entire bundle* of resources across persons, that is, would count as appropriately compensating individuals for their endowment of non-transferable resources? Dworkin's answer was to construct a kind of market for contingent claims behind a thin veil of ignorance in which traders knew their preferences (importantly, over risk) but not what resources they would come to have in the (actual) world. The desirable tax scheme, in the world, would mimic the allocation of transferable resources that would be implemented at the

equilibrium in this market for contingent claims, after the birth lottery occurred (see Roemer 1996, ch. 7, for a formal model).

Dworkin's contribution, importantly, attempted to integrate the issue of responsibility into egalitarian theory – which amounted to taking the most important tool of the political right and harnessing it for use by the political left. In Dworkin's theory, individuals are held responsible for their preferences, and this is implemented through the insurance market behind the veil of ignorance, where traders representing persons use their persons' preferences to enter into insurance contracts. But persons are not held responsible for their resources, including internal talents, and the families into which they are born, and this is implemented through allowing the traders behind the veil to insure against bad luck in the birth lottery in so far as the distribution of these resources is concerned.

Several years later, G. A. Cohen (1989) and Richard Arneson (1989) criticized Dworkin's theory. Cohen argued that 'Dworkin's cut' between preferences and resources, was, for the purpose of ethics, the wrong way to separate characteristics. Suppose a person developed champagne tastes because she grew up in an aristocratic family in which she was never exposed to beer. Was it correct to later deny her the resources to buy champagne to achieve the level of welfare that beer drinkers could achieve more cheaply? Or suppose that a person, who grew up in a disadvantaged home that lacked resources, developed no ambition to develop his talents; indeed, he was satisfied with his unambitious tastes. Should he likewise be held responsible, even though his tastes were the consequence, at least in part, of an indigent childhood?

Arneson argued that Dworkin was right to argue against taking the 'equalisandum' as welfare, but said that replacing it with 'resources' was wrong – rather, it should be replaced with *opportunity for welfare.* What did it mean, then, to equalize opportunities for welfare? In what sense did this differ from 'equalizing resources' à la Dworkin? Arneson struggled to formulate an alternative, but did not succeed in proposing one that was clearly feasible.

Following Arneson and Cohen, Roemer (1993, 1998) proposed a model that would attempt to capture the insights of this philosophical discussion and permit one to compute, for a given situation, the policy that constituted the 'equal opportunity' policy. He separated the influences on the outcome a person experiences into *circumstances* and *effort:* circumstances are attributes of the person's environment for which he should not be held responsible, and effort is the choice variable for which he should be held responsible. An equal-opportunity (EOp) policy is an intervention (such as the provision of resources by a state agency) that makes it the case that all those who expend the same degree of effort end up with the same outcome, regardless of their circumstances. Thus, EOp 'levels the playing field' in the sense of compensating persons for their deficits in circumstances, ensuring that, finally, only effort counts with regard to outcome achievement.

A more precise formulation follows. Suppose there is an *objective* for whose acquisition a planner wishes to equalize opportunities; this might be a wage-earning capacity, a life expectancy, or an income level. Denote the achievement of the objective as a function $u(\alpha, x; \beta)$ where α is the (scalar) level of effort expended by the person, x is the policy of the planner, and β is the vector of circumstances of the person. u is monotone increasing in the argument α – thus, effort enhances the acquisition of the objective. Nevertheless, effort may be subjectively costly for the individual: thus, u is not to be thought of as the usual economist's utility function, in which effort is costly. For example, u might be the wage-earning capacity a person comes to have, where α is the number of years of schooling and β measures family background, natural talent, and so on. The policy x can be chosen from some feasible set of policies X: it might, for example, be the distribution of a resource possessed by the planner. The set of individuals with a given value of β is called a 'type'.

Suppose that, for each ordered pair (x, β) there ensues a distribution of effort in type β denoted by its cumulative distribution function $F(\alpha; x, \beta)$. The distribution of effort, classically, would result from the maximization of a preference order by

the individuals of the type, one in which effort is differentially costly for those individuals. Typically, these distributions F differ across types (β). By hypothesis, individuals are not to be held responsible for their type. We now ask how one should interpret the stricture to choose a policy that equalizes the values of the objective *at constant effort levels* across types. The problem is that the *distribution* of effort in a type is a characteristic of the type and, if individuals are to be compensated for their types, they should likewise be compensated for the characteristics of those distributions. For example, if a disadvantaged type has a distribution of effort with a low mean, that itself should be taken into account in the compensation scheme. Roemer's solution was to propose that the *degree* of a person's effort should be measured by her *rank* in the effort distribution of her type. Thus, define the rank π by

$$\pi = F(\alpha; x, \beta)$$

and define the ' indirect' objective function

$$v(\pi; x, \beta) = u\big(F^{-1}(\pi; x, \beta), x; \beta\big).$$

Then x is an equal-opportunity policy just in case it equalizes the value of objective across types at every degree of effort, that is:

$$\forall \pi \in [0,\ \ 1] \forall \beta,\ \ \beta' v(\pi; x, \beta) = v(\pi; x, \beta'). \quad (1)$$

Here, the process by which the effort distributions F emerge is black-boxed; of course, in actual applications, the black box would be unpacked with the specification of utility functions that individuals maximize to derive their efforts.

In general there will be no policy which equalizes opportunities in the sense of (1). For example, let u be wage-earning capacity, α be years of schooling, β be the educational level of the individual's parents, and x be investment in the education of the individual by the state. Suppose policies can be targeted to types, and there is a per capita social endowment of $\bar{x}$ for education. Suppose we partition the population into a finite set of types, $\{\beta_i |\ i = 1,\ \ \ldots,\ n\}$ where the population frequency of type i is p_i. A feasible policy is a vector $(x_1,\ \ \ldots,\ x_n)$ such that $\Sigma p_i x_i = \bar{x}$. We have as data, as well, the distribution functions $F(\cdot; x,\ \ \beta)$. For this general specification, there will generally not exist a feasible policy satisfying (1).

Some alternative is therefore required. One may proceed as follows. We desire to equalize the values of v across different β's, at each π. As a second-best, we desire to maximize the minimum value of v across different β's, at each π. Thus, define

$$\Phi(\pi; x) = \min_{\beta}\ \ v(\pi, x, \beta).$$

We define a policy x to be *efficient* if

$$\text{there is no } \ x' \in X \quad s.t(\forall \pi)(\Phi(\pi; x') \geq \Phi(\pi; x)), \quad (2)$$

where the inequality sign in (2) is understood to mean that, for some value(s) of π, there is strict inequality. We are interested only in efficient policies. There may, however, be many, even a continuum, of these; and the theory, thus far, gives us no way of choosing among them.

To see this, let us consider a special case in which effort responses within types are insensitive to the policy: thus, we may write those distributions as $F(\alpha;\ \beta)$. Suppose that there are just two types, $\beta = 1$ and $\beta = 2$, indicating the level of education of parents; each type comprises one-half the population. The Department of Education has one unit per capita of an educational resource to be invested in children. A *policy* is an ordered pair $(y,\ 2-y)$, indicating the per capita investment in children of the two types. Suppose that $u(\alpha,\ y;\ \beta) = \alpha^a\ y^c\ \beta^b$ is the value of the objective (perhaps, the child's future wage) where y is the amount of educational resource invested in the child. We will denote a policy by the value of its first component, y. Then

$$\Phi(\pi; y) = \ \ \min_{y}\ \ \left[\left(F^{-1}(\pi; 1)\right)^a y^c, \left(F^{-1}(\pi; 2)^a 2^b (2-y)^c\right].\right. \quad (3)$$

We may compute that the two arguments of the min function in (3) are equalized exactly when

$$y = \frac{2}{1 + \left(\frac{1}{2}\right)^{b/c}\left(\frac{F^{-1}(\pi;1)}{F^{-1}(\pi;2)}\right)^{a/c}}. \tag{4}$$

Now define

$$m = \min_{\pi}\left(\frac{F^{-1}(\pi;1)}{F^{-1}(\pi;2)}\right), M = \max_{\pi}\left(\frac{F^{-1}(\pi;1)}{F^{-1}(\pi;2)}\right). \tag{5}$$

Then any policy

$$y \in \left[\frac{2}{1 + \left(\frac{1}{2}\right)^{b/c} M^{a/c}}, \frac{2}{1 + \left(\frac{1}{2}\right)^{b/c} m^{a/c}}\right] \tag{6}$$

is efficient, and this interval comprises exactly the efficient policies. Thus, there is a continuum of efficient policies.

There has been no general agreement concerning how to narrow the set of efficient policies to a single choice – in other words, how to rank efficient policies from the equal-opportunity viewpoint. Roemer (1998) proposed to choose a single policy by solving the problem:

$$\max_{x} \int_0^1 \Phi(\pi; x) d\pi; \tag{7}$$

Van de Gaer (1993) proposed to solve

$$\max_{x} \min_{\beta} \int v(\pi; x, \beta) d\pi. \tag{8}$$

Each of these proposals is somewhat arbitrary. Fleurbaey and Maniquet (2004) summarize the axiomatic approach to the problem, to which they and others have made substantial contributions. I believe that appeal to the equal-opportunity principle as such cannot resolve the issue; we must bring additional ethical considerations to bear.

How does the theory of equal opportunity fit into social choice theory? There are a number of ways one may answer this question; I believe the most salient point is that the equal-opportunity approach is distinguished from classical social-choice theory in being *non-welfarist.* Welfarism is the view that only the set of vectors of outcome (welfare) possibilities matters for the social decision. To be precise, if we represent individual preferences over social alternatives by utility functions, then the choice of a social alternative should depend only upon the information that is recoverable from the utility possibilities sets of the possible societies. In this sense, welfarism is a consequentialist view. Sen (1979) criticized the welfarist postulate for ignoring the issue of civil rights (the right not to be beaten by another, for instance); Roemer (1996) criticized it, with regard to the theory of distributive justice, for ruling out any theories which mention property rights. The equal-opportunity approach says that one cannot judge the goodness of a social outcome by knowing only the distribution of outcomes; one must also know *how hard people tried* in order to evaluate that goodness – in other words, one must know the correlation of effort with achievement to pass judgement on the fairness of a distribution scheme. Put this way, it is clear that the equal-opportunity approach formalizes a view that is held quite generally by citizens in many countries, to judge from opinion surveys. In judging how just schemes of distribution are, the proverbial man on the street usually wants to know if reward is 'proportionate' to effort expended. Knowing only the distribution of outcomes does not suffice.

Several empirical studies have applied these ideas. In Roemer et al. (2003), the authors asked: in a set of 11 OECD countries, what income-tax regime would equalize opportunities for income acquisition among workers? All workers in a country were assumed to have a quasi-linear utility function over income and labour, with a constant labour-supply elasticity with respect to the marginal tax rate (or the wage). The sole circumstance was taken to be the level of education of the mother of the worker. Young male workers were partitioned into three types, according to whether their mothers had low, medium, or high levels of education. The set of policies, X, was taken to be the set of feasible affine income tax regimes, that is, ones postulating constant marginal tax rates

Equality of Opportunity, Table 1 EOp marginal income tax rates for 11 countries

Country	Observed marginal income-tax rate	EOp marginal income tax rate
Belgium	.53	.54
Denmark	.44	0
France	.31	.58
Great Britain	.36	.71
Italy	.23	.82
Netherlands	.53	.47
Norway	.39	0
Spain	.38	.61
Sweden	.52	0
United States	.24	.65
West Germany	.36	0

Source: Roemer et al. (2003)

and a lump-sum payment to all. The objective was the post-fisc income (not utility) of the individual. Using the EOp objective of (7) turns out to be equivalent to choosing that income-tax regime which maximizes the average post-fisc income of the least advantaged type, those whose mothers did not complete secondary school. Table 1 summarizes the observed marginal tax rates in the countries of the sample and the equal-opportunity tax rates, so computed, under the assumption that the (male) labour-supply elasticity with respect to taxation is – .06.

Countries can be partitioned into three groups: those for which observed tax rates are much greater than the EOp tax rate (West Germany, Denmark, Sweden and Norway), those for which the observed and EOp tax rates are approximately the same (Belgium and the Netherlands), and those for which observed tax rates are much lower than the EOp tax rates (Italy, Spain, France, the United States and Great Britain). The pattern is not particularly surprising given common perceptions of the depth of income-transfer programmes in these countries.

A comment upon the countries in the first category is in order. To say that the EOp tax rate is zero in the northern European countries means that, with the postulated labour-supply effects of taxation, the average post-fisc income of the least advantaged type would be maximized with a lump-sum tax to finance public goods, and no other transfer payments. This comes about precisely because the *pre-fisc* distributions of income across the three types of worker are already very close in these countries. In the other countries in the sample, these pre-fisc distributions are sufficiently far apart that positive marginal tax rates will, despite their incentive effects, increase the average post-fisc income of the least advantaged type.

Should one conclude from Table 1 that, from the equal-opportunity viewpoint, marginal income taxation should be abandoned in northern Europe? Hardly, for the division of workers into only three types is quite coarse. There are many other circumstances besides the education of the mother for which society might wish to compensate citizens. Indeed, the article under discussion studies as well a typology for four of the countries (where data exist) into six types, where workers are typed not only by three maternal education levels but also by two levels of native ability, as measured by performance in IQ tests. It turns out that a positive marginal EOp tax rate is then recommended for Sweden, although Denmark retains its zero tax rate! (With a sufficiently low labour-supply elasticity, this result, too, would be changed.)

Income taxation may not be the instrument of choice to equalize opportunities for income; one naturally thinks of using educational finance policy as a method for compensating children from disadvantaged families. Betts and Roemer (2003) partitioned American male workers who were attending secondary school in the late 1960s into four types, defined by four levels of maternal education. They took *wage-earning capacity* as the objective and state educational investment in the child as the policy instrument, and asked: What distribution of educational finance would have equalized opportunities for wage-earning capacity among these four types of worker? Wage elasticities with respect to educational investment were computed for the four types using data from the US Panel Studies on Income Dynamics (PSID). Based on the assumption of a per capita educational budget of $2500, the recommended allocation is presented in Table 2.

Equality of Opportunity, Table 2 EOp allocation of investment with per capita budget of $2500 per student per annum

Parental ed'n	< 8 years	8 < ed < 12 years	12 years	> 12 years
EOp investment	$5360	$3620	$1880	$1100

Source: Betts and Roemer (2003)

Equality of Opportunity, Table 3 EOp allocation of educational investment, four types, race × maternal education

Type of worker	LB	HB	LW	HW
EOp investment	$8840	$16,260	$2610	$679

Source: Betts and Roemer (2003)

In other words, equal-opportunity investment would allocate almost five times as much to the most disadvantaged type of student as to the most advantaged type. Interestingly, we computed that the average wage of workers under this allocation would have risen by 2.6 per cent over the observed average wage. In other words, there is no observed trade-off between equity and 'efficiency'.

The authors computed that if the allocation of Table 2 had been implemented there would have been very little change in the fraction of black workers who would have risen above the bottom quintile of the wage distribution. They proceeded to compute the EOp policy for a different typology of workers into four types, defined as:

LB: low maternal education, black
HB: high maternal education, black
LW: low maternal education, white
HW: high maternal education, white

The results are presented in Table 3.

For this typology, the investment ratios are huge. Moreover, the total wage bill would fall by 2 per cent under the allocation of Table 3, showing that an equity-efficiency' trade-off does exist with respect to this typology.

At the least, the calculations of Betts and Roemer demonstrate that there is a large difference between an *equal-resource policy,* which invests the same amount in all children, and an *equal-opportunity policy,* which invests in children in such a way as to attempt compensation for differential social circumstances. The United States, with its system of locally financed public education, is in most places less equitable even than the equal-resource policy would be: that is, usually more is invested in the public education of advantaged children than in that of disadvantaged children.

I have earlier distinguished between the equal-opportunity approach and the more classical welfarist approach in welfare economics. A second important distinction is between equal opportunity, as a concept of equity, and meritocracy. Consider the problem of admissions to university or professional school. The equal-opportunity approach would suggest admitting the highest-effort candidates from each of a set of types, distinguished by their levels of advantage in background. The meritocratic approach would suggest admitting those who are most likely to be high achievers. EOp focuses upon fair treatment *among the pool of candidates,* while meritocracy has a double focus: treating the candidates fairly but also considering the quality of services those candidates will, in the future, provide to society at large. (On the other hand, meritocracy is *not* concerned with candidates' effort in its measurement of fair treatment, but only with their ability to perform.) Thus the two approaches are in conflict.

Clearly, the quality of services provided to society at large must count – the unadorned EOp approach cannot in general be the right one. Generally speaking, society should follow a mixture of equal-opportunity and meritocratic policies. To calibrate the right mixture would require, as well as data to calculate the relevant elasticities, a general theory of justice for society at large, in which account is taken not only of fairness to those competing to occupy social positions but

of the welfare of those who eventually consume the products those individuals will produce. In the US debate around affirmative action, one can hear different emphases. With respect to school admissions, most citizens seem concerned with fairness to the candidates, although there is a dispute as to what traits should or should not count in judging fairness; but, with respect to employment, many believe meritocratic principles are primary. Thus, race-based affirmative action policies in universities are under challenge for focusing on the wrong parameters of disadvantage (which, many argue, should be ones of social class, not race), while affirmative-action employment policies are challenged for paying insufficient attention to competence in employing workers.

In the applications discussed above, the policymakers – whether fictitious ones in the minds of scholars or actual ones in social institutions – have generally contemplated only the effects of policies in a single sector, whether it be in education or employment. Calsamiglia (2005) has posed the following problem. Suppose individuals are competing for positions in several sectors simultaneously (in her example, for admission to a university and to an athletic team), and the admissions officer in each sector is attempting to design an equal-opportunity policy for the candidates in his sector alone. Thus, the university admissions officer knows the abilities and circumstances and efforts of candidates for university, and the athletic coach knows the same information as it applies to performance in her sector. Each designs a *local* equal-opportunity admissions policy for his own sector. When will the combination of policies equalize opportunities *globally?* The tension here is that policies in each sector will, if not properly designed, distort the efforts of candidates in other sectors. Calsamiglia demonstrates that, under suitable conditions, locally designed EOp policies aggregate into a global EOp policy if and only if they *equalize rewards to effort* across types in each sector. For example, assigning disadvantaged students who are applying to law school 'extra points' to compensate them does *not* equalize rewards to effort as between them and more advantaged students: rather, one requires a policy which, for each unit of effort expended, increases the probability of admission by the *same amount* across all types of student. One can say, that is, that equalizing rewards to effort is the necessary and sufficient condition for decentralizing the social problem of equalizing opportunities across the board into policy formation at the sectoral level. Whether or not Calsamiglia's insight will be important in policy design depends upon the degree to which individuals are involved in inter-sectoral effort allocation decisions.

See Also

- ▶ Justice
- ▶ Redistribution of Income and Wealth

Bibliography

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56: 77–93.

Betts, J., and J. Roemer. 2003. Equalizing opportunity through reform. In *Schools and the equal opportunity problem*, ed. P. Peterson and L. Woessmann. Cambridge, MA: MIT Press.

Calsamiglia, C. 2005. Decentralizing equality of opportunity and issues concerning the equality of educational opportunity. Ph.D. dissertation. New Haven: Department of Economics, Yale University.

Cohen, G. 1989. On the currency of egalitarian justice. *Ethics* 99: 906–944.

Dworkin, R. 1981a. What is equality? Part one: Equality of welfare. *Philosophy & Public Affairs* 10: 185–246.

Dworkin, R. 1981b. What is equality? Part two: Equality of resources. *Philosophy & Public Affairs* 10: 283–345.

Fleurbaey, M., and F. Maniquet. 2004. Compensation and responsibility. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura, vol. 2. Amsterdam: North-Holland.

Roemer, J. 1993. A pragmatic theory of responsibility for the egalitarian planner. *Philosophy & Public Affairs* 22: 146–166.

Roemer, J. 1996. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.

Roemer, J. 1998. *Equality of opportunity.* Cambridge, MA: Harvard University Press.

Roemer, J., R. Aaberge, U. Colombino, J. Fritzell, S. Jenkins, I. Marx, M. Page, E. Pommer, J. Ruiz-Castillo, M.J. San Segundo, T. Tranaes, G. Wagner, and I. Zubiri. 2003. To what extent do fiscal regimes equalize opportunities for income acquisition among citizens? *Journal of Public Economics* 87: 539–565.

Sen, A. 1979. Utilitarianism and welfarism. *Journal of Philosophy* 76: 463–489.
Van de Gaer, D. 1993. Equality of opportunity and investment in human capital. Ph.D. dissertation. Leuven: Catholic University of Leuven.

Equation of Exchange

Michael D. Bordo

Abstract

One of the oldest formal relationships in economics, the equation of exchange, as a basic accounting identity of a money economy, demonstrates that the sum of expenditures must equal the sum of receipts. It is useful both as a classification scheme for analysing the underlying forces at work in a money economy and as a building block or engine of analysis for monetary theory and in particular for the quantity theory of money. It can also be regarded as a building block for a macro theory of aggregate demand and supply, and used to construct a theory of nominal income.

Keywords

Briscoe, J.; Cairnes, J. E.; Cash balance (Cambridge) approach; Circular flow of income; De Foville, A.; Equation of exchange; Equation of societary circulation; Fisher, I.; Friedman, M.; Hadley, A. T.; Kemmerer, E. W.; Keynes, J. M.; Lang, J.; Law, J.; Levasseur, E.; Lloyd, H.; Lubbock, J.; Mill, J. S.; Naive quantity theory; Newcomb, S.; Nominal income; Norton, J. P.; Pantaleoni, M.; Pigou, A. C.; Quantity equation; Quantity theory of money; Rau, K. H.; Ricardo, D.; Senior, N. W.; Stocks and flows; Thornton, H.; Transactions velocity of circulation; Turner, S.; Velocity of circulation; Walras, L.

JEL Classifications

E4

The equation of exchange (often referred to as the quantity equation) is one of the oldest formal relationships in economics, early versions of both verbal and algebraic forms appearing at least in the 17th century. Perhaps the best known variant of the equation of exchange is that expressed by Irving Fisher (1922):

$$MV = PT. \quad (1)$$

Equation 1 represents a simple accounting identity for a money economy. It relates the circular flow of money in a given economy over a specified period of time to the circular flow of goods. The left-hand side of Eq. 1 stands for money exchanged, the right-hand side represents the goods, services and securities exchanged for money during a specified period of time. M is defined as the total quantity of money in the economy, T as the total physical volume of transactions, where a transaction is defined as any exchange of goods, including physical capital, services and securities for money, P is an appropriate price index representing a weighted average of the prices of all transactions in the economy. Finally, to make the stock of money comparable with the flow of the value of transactions (PT), and to make the two sides of the equation balance, it is multiplied by V, the transactions velocity of circulation, defined as the average number of times a unit of currency turns over (or changes hands) in the course of effecting a given year's transactions.

An alternative variant of the equation of exchange is the income version by Pigou (1927). Empirical difficulties in measuring an index of transactions, and the special price index related to it, led, with the development of national income accounting, to the formulation of Eq. 2:

$$MV = PY \quad (2)$$

where Y represents national income expressed in constant dollars, P the implicit price deflator and V the income velocity of circulation defined as the average number of times a unit of currency turns over in the course of financing the year's final activity.

Equations 1 and 2 differ from each other because the volume of transactions in the economy includes intermediate goods and the exchange of existing assets, in addition to final goods and services. Thus vertical integration and other factors which affect the ratio of transactions to income would also alter the ratio of transactions velocity to income velocity.

A third version of the equation of exchange, the Cambridge cash balance approach (Pigou 1917; Marshall 1923; Keynes 1923), converts the flow of spending into units comparable to the stock of money

$$M = kPY \tag{3}$$

where $k = 1/V$ is defined as the time duration of the flows of goods and services money could purchase, for example, the average number of weeks income held in the form of money balances.

Equations 2 and 3 are arithmetically equivalent to each other but they rest on fundamentally different notions of the role of money in the economy. Both Eqs. 2 and 1 view money primarily as a medium of exchange and the quantity of money is represented as continually 'in motion' – constantly changing hands from buyer to seller in the course of a time period. Equation 3 views money as a temporary abode of purchasing power (an asset) forming part of a cash balance 'at rest'. Consequently, the items included in the definition of money in the transactions and income versions of the equation of exchange are assets used primarily to effect exchange – currency and checkable deposits, whereas the cash balance approach includes, in addition to these items, non-checkable deposits and possibly other liquid assets.

The equation of exchange is useful both as a classification scheme for analysing the underlying forces at work in a money economy and as a building block or engine of analysis for monetary theory and in particular for the quantity theory of money.

As a classification scheme, the equation as a basic accounting identity of a money economy demonstrates the two-sided nature of the circular flow of income – that the sum of expenditures must equal the sum of receipts. The left-hand side of the equation shows the market value of goods and services purchased (dollar value of goods exchanged) and the money received. The equation also relates the stock of money to the circular flow of income by multiplying M by its velocity. Finally, the equation is useful in creating definitional categories – M, V, P, T – amenable both to empirical measurement and to theoretical analysis.

The equation of exchange is best known as a building block for the quantity theory of money. The traditional approach has been to make behavioural assumptions about each of the variables in the equation, converting it from an identity to a theory. The simplest application, dubbed the 'naive quantity theory' (Locke 1691) treated V and T in Eq. 1 as constants, with P varying in direct proportion to M.

A more sophisticated version (Fisher 1911) treats each of M, V and T as being normally determined by independent sets of forces, with V as determined by slowly changing factors such as those affecting the payments process and the community's money holding habits.

The Cambridge cash balance approach, based on Eq. 3, views the quantity theory as encompassing both a theory of money demand and money supply. In this approach the nominal money supply is determined by the monetary standard and the banking system while the nominal quantity of money demanded is proportional to nominal income, with k the factor of proportionality, representing the community's desired holding of real cash balances. k in turn is determined by economic variables such as the rate of interest in addition to the factors stressed by the Fisher approach. The price level (value of money) is then determined by the equality of money supply and demand.

The equation of exchange can also be regarded as a building block for a macro theory of aggregate demand and supply (Schumpeter 1954). If we view MV as aggregate demand and T or Y as aggregate supply, then P would be determined in the familiar Marshallian way.

Finally, the equation can be used to construct a theory of nominal income. According to this

approach (Friedman and Schwartz 1982), nominal income is determined by the interaction of the money supply and a stable demand for real cash balances. The decomposition of a given change in nominal income into a change in the price level and in real output is determined in the short run by inflation (deflation) forecast errors and in the long run by the natural rate of output.

The equation of exchange both as a classification scheme and as a building block for the quantity theory of money can be traced back to the earliest development of economic science.

The pre-classical writers of the 17th and 18th centuries viewed the equation in both senses. Locke (1691), Hume (1752) and Cantillon (1735) each organized his approach to monetary issues using the equation. Locke had a clear statement of the naive quantity theory assuming both V and T to be immutable constants. Hume followed Locke but made a clear distinction between long-run statics and short-run dynamics. In the long run the price level would be proportional to M but, in the short run or transition period, changes in M would produce changes in T. Cantillon had a clear understanding of the relationship between the stock of money and the circular flow of income. Indeed, he was the first to define explicitly the concept of velocity of circulation, viewing V not as a constant but as a variable influenced in a stable way by both technological and economic variables. Furthermore, like Hume, Cantillon distinguished between the long-run equilibrium nature of the quantity theory and short-run disequilibrium. Both Locke and Hume viewed the equation from the perspective of money 'at rest' forming a cash balance whereas Cantillon viewed money as continuously in 'motion'.

John Law (1705) understood the equation of exchange but used it to derive a link between changes in the quantity of M and changes in T.

The classical economists Thornton,et al. followed the Locke/Hume/Cantillon tradition of the quantity theory of money using a verbal version of the equation of exchange in their monetary analysis.

Algebraic versions of the equation first appeared in the 17th and 18th centuries (see Marget 1942; Humphrey 1984). The British writers Briscoe (1694) and Lloyd (1771) both expressed a rudimentary version of Eq. 1, unfortunately omitting a term for velocity. Turner (1819) formulated the equation without breaking PT into separate components. The most complete early statement of the equation was by Sir John Lubbock (1840), who not only included all the items of the equation but (preceding Fisher) distinguished between the quantities and velocities of hard currency, bank notes and bills of exchange. Similar complete algebraic statements of the equation were made by the German writers Lang (1811) and Rau (1841); the Italian Pantaleoni (1889); the Frenchmen Levasseur (1858), Walras (1874) and de Foville (1907); and the Americans Newcomb (1885), Hadley (1896), Norton (1902) and Kemmerer (1907). Of this group Newcomb presented the clearest statement. Newcomb started with the concept of exchange as involving the transfer of money for wealth. Summing up all exchanges in the economy he arrived at his equation of societary circulation:

$$VR = KP \tag{4}$$

where V represents the total value of currency, R the rapidity (velocity) of circulation, K the volume of real transactions, P a price index.

The clearest and best known algebraic expressions of the equation were by the neoclassical economists Irving Fisher (1922) and A.C. Pigou (1917). Fisher (1911, pp. 15–17), directly following Newcomb, defined the equation of exchange as

> a statement, in mathematical form, of the total transaction: effected in a certain period in a given community. . . . [I]n the grand total of all exchanges for a year, the total money paid is equal to the total value of goods bought. The equation thus has a money side and a goods side. The money side is the total money paid, and may be considered as the product of the quantity of money multiplied by its rapidity of circulation. The goods side is made up of the products of quantities of goods exchanged multiplied by their respective prices.

This statement expressed as in Eq. 1 or in an expanded version distinguishing between currency and deposits payable by check,

$$MV + M'V' = PT \tag{5}$$

where M' is defined as checkable deposits and V' their velocity, Fisher then used to analyse the forces determining the price level.

Fisher's approach followed the 'motion' theory tradition of Cantillon with velocity determined primarily by technological and institutional factors. In contrast, Pigou (1917) and other writers in the Cambridge tradition, Marshall (1923) and Keynes (1923), followed the 'rest' approach of Locke and Hume, expressing the equation as

$$1/P = kR/M \tag{6}$$

where R represents total resources enjoyed by the community, k the proportion of resources the community chooses to keep in the form of titles to legal tender, M the number of units of legal tender and P a price index. For Pigou the fundamental difference between his approach and that of Fisher was that by focusing

> attention on the proportion of their resources that people *choose* to keep in the form of titles to legal tender instead of focusing on the 'velocity of circulation' . . . it brings us . . . into relation with *volition* – an ultimate *cause of demand* – instead of with something that seems at first sight *accidental and arbitrary*. (1917, p. 174, emphasis added)

The Cambridge cash balance version of the equation of exchange, by focusing on the demand for money and volition rather than emphasizing mechanical aspects of the circular flow of money, can be viewed as the starting point for the Keynesian approach to the demand for money (Keynes 1936), for modern choice theoretic approaches to money demand (Hicks 1935) and for the modern quantity theory of money (Friedman 1956).

See Also

- ▶ Newcomb, Simon (1835–1909)
- ▶ Quantity Theory of Money

Bibliography

Bordo, M.D. 1983. Some aspects of the monetary economics of Richard Cantillon. *Journal of Monetary Economics* 12: 234–258.

Briscoe, J. 1694. *Discourse on the late funds* .... London.

Cantillon, R. 1755. In *Essai sur la nature du commerce en général*, ed. H. Higgs. London: Macmillan, 1931; reprinted New York: Augustus M. Kelley, 1964.

Fisher, I. 1911. *The purchasing power of money*. New York: Macmillan.

Fisher, I. 1922. *The purchasing power of money*, 2nd ed. Reprinted New York: Augustus M. Kelley, 1963.

Foville, A. de. 1907. *La monnaie*. Paris.

Friedman, M. 1956. The quantity theory of money – A restatement. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M., and A.J. Schwartz. 1982. *Monetary trends in the United States and the United Kingdom: Their relation to income, prices and interest rates, 1867–1975*. Chicago: University of Chicago Press for the NBER.

Hadley, A.T. 1896. *Economics*. New York.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* 2: 1–19.

Holtrop, M.W. 1929. Theories of the velocity of circulation of money in earlier economic literature. *Economic Journal* 39 (January): 503–524.

Hume, D. 1752. Of money. In *Essays, moral, political and literary*, vol. 1 of *Essays and treatises*, a new edition, Edinburgh: Bell and Bradfute; Cadell and Davies, 1804.

Humphrey, T.M. 1984. Algebraic quantity equations before Fisher and Pigou. *Federal Reserve Bank of Richmond Economic Review* 70 (5): 13–22.

Kemmerer, E.W. 1907. *Money and credit instruments in their relation to general prices*. New York: H. Holt & Co.

Keynes, J.M. 1923. *A tract on monetary reform*. Reprinted, London: Macmillan for the Royal Economic Society, 1971.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. Reprinted, London: Macmillan for the Royal Economic Society, 1973.

Lang, J. 1811. *Grundlinien der politischen Arithmetik*. Kharkov.

Law, J. 1705. *Money and trade considered with a proposal for supplying the nation with money*. New York: Augustus Kelley, 1966.

Levasseur, E. 1858. *La question de l'or: les mines de Californie et d'Australie*. Paris.

Lloyd, H. 1771. *An essay on the theory of money*. London.

Locke, J. 1691. *The works of John Locke*, Vol. 5. London, 1823.

Lubbock, J. 1840. *On currency*. London.

Marget, A.W. 1942. *The theory of prices*. New York: Prentice-Hall.

Marshall, A. 1923. *Money, credit and commerce*. London: Macmillan. Reprinted, New York: Augustus M. Kelley, 1965.

Newcomb, S. 1885. *Principles of political economy*. New York: Harper & Brothers.

Norton, J.P. 1902. *Statistical studies in the New York money market*. New York: Macmillan.
Pantaleoni, M. 1889. *Pure economics*. Trans. T.B. Bruce, London: Macmillan, 1898.
Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32 (November), 38–65. Reprinted in *Readings in Monetary Theory*, Edited by F.A. Lutz and L.W. Mints for the American Economic Association, Homewood: Irwin, 1951.
Pigou, A.C. 1927. *Industrial fluctuations*, 2nd ed. London: Macmillan, 1929.
Rau, K.H. 1841. *Grundsatze der Volkswirtsaftslehre*, 4th ed. Leipzig and Heidelberg.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Turner, S. 1819. *A letter addressed to the right Hon. Robert Peel with reference to the expediency of the resumption of cash payments at the period fixed by law*. London.
Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: Corbaz.

Equilibrium (Development of the Concept)

Murray Milgate

Keywords

Ceteris paribus; Equilibrium; Expectations; Game theory; General equilibrium; Intertemporal equilibrium; Long-run equilibrium; Natural and normal conditions; Partial equilibrium; Short-run equilibrium; Stationary state; Steuart, J.; Temporary equilibrium

JEL Classifications

B0

From what appears to have been the first use of the term in economics by James Steuart in 1769, down to the present day, equilibrium analysis (together with its derivative, disequilibrium analysis) has been the foundation upon which economic theory has been able to build up its not inconsiderable claims to 'scientific' status. Yet despite the persistent use of the concept by economists for over 200 years, its meaning and role have undergone some quite profound modifications over that period.

At the most elementary level, 'equilibrium' is spoken about in a number of ways. It may be regarded as a 'balance of forces', as when, for example, it is used to describe the familiar idea of a balance between the forces of demand and supply. Or it can be taken to signify a point from which there is no endogenous 'tendency to change': stationary or steady states exhibit this kind of property. However, it may also be thought of as that outcome which any given economic process might be said to be 'tending towards', as in the idea that competitive processes tend to produce determinate outcomes. It is in this last guise that the concept seems first to have been applied in economic theory. Equilibrium is, as Adam Smith might have put it (though he did not use the term), the centre of gravitation of the economic system – it is that configuration of values towards which all economic magnitudes are continually tending to conform.

There are two properties embodied in this original concept which when taken into account begin to impart to it a rather more precise meaning and a well-defined methodological status. Into this category enters the formal definition of 'equilibrium conditions' and the argument for taking these to be a useful object of analysis.

There are few better or more appropriate places to isolate the first two properties of 'equilibrium' in this original sense than in the seventh chapter of the first book of Adam Smith's *Wealth of Nations.* The argument there consists of two steps. The first is to define 'natural conditions':

> 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 wages of the labour and the profits of the stock employed . . . according to their natural rates, the commodity is then sold for what may be called its natural price. (Smith 1776, I.vii, p. 62)

The key point here is that 'natural conditions' are associated with a general rate of profit – that is, uniformity in the returns to capital invested in different lines of production under existing best-practice technique. In the language of the day, this

property was thought to be the characteristic of the outcome of the operation of the process of 'free competition'.

The second step in the argument captures the analytical status to be assigned to 'natural conditions':

> 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 even somewhat below it. But whatever may be the obstacles which hinder them from settling in this center of repose and continuance, they are constantly tending towards it. (I.vii, p. 65)

This particular 'tendency towards equilibrium' was held to be operative in the *actual* economic system at any given time. It is not to be confused with the familiar question concerning the stability of competitive equilibrium in modern analysis. There the question about convergence to equilibrium is posed in some *hypothetical* state of the world where none but the most purely competitive environment is held to prevail. It is also essential to observe that in defining 'natural conditions' in this fashion, nothing has yet been said (nor need it be said) about the forces which act to determine the natural rates of wages and profits, or the natural prices of commodities. It will therefore be possible to refrain from discussing the *theories* offered by various economists for the determination of these variables in most of what follows. Treatment of these matters may be found elsewhere in this dictionary. Similarly, there will be no discussion here of existence or uniqueness of equilibrium (see existence of general equilibrium).

'Natural conditions' so defined and conceived are the formal expression of the idea that certain systematic or persistent forces, regular in their operation, are at work in the economic system. Smith's earlier idea, that 'the co-existent parts of the universe . . . contribute to compose one immense and connected system' (1759, VII.ii, 1.37), is translated in this later formulation into an analytical device capable of generating conclusions with a claim to general (as opposed to a particular, or special) validity. These general conclusions were customarily referred to as 'statements of tendency', or 'laws', or 'principles' in the economic literature of the 18th and 19th centuries. It is worth emphasizing that there was no implication that these general tendencies were either swift in their operation or that they were not subject at any time to interference from other obstacles. Like sea level, 'natural conditions' had an unambiguous meaning, even if subject to innumerable cross-currents.

To put it another way, the distinction between 'general' and 'special' cases (like its counterpart, the distinction between 'equilibrium' and 'disequilibrium'), refers neither to the immediate practical relevance of these kinds of cases to actual existing market conditions, nor to the prevalence, frequency, or probability of their occurrence. In fact, as far as simple observation is concerned, it might well be that 'special' cases would be the order of the day. John Stuart Mill expressed this idea especially clearly when he held that the conclusions of economic theory are only applicable 'in the *abstract*', that is, 'they are only true under certain suppositions, in which none but general causes – causes common to the *whole class* of cases under consideration – are taken into account' (Mill 1844, pp. 144–5). Marshall, of course, understood their application as being subject not only to this qualification (which he spoke about in terms of 'time'), but also to the condition that 'other things are equal' (1890, I.iii, p. 36). There will be cause to return to this matter below.

To unearth these regularities, one had to inquire behind the scene, so to speak, to reveal what otherwise might remain hidden. Adam Smith had set out the basis of this procedure in an early essay on 'The Principles which Lead and Direct Philosophical Enquiries':

> Nature, after the largest experience that common observation can acquire, seems to abound with events which appear solitary and incoherent. . . by representing the invisible chains which bind together all these disjointed objects, [philosophy] endeavours to introduce order into this chaos of jarring and discordent appearances. (Smith 1795, p. 45)

In short, 'equilibrium', if we may revert to the modern terminology for a moment, became the

central organizing category around which economic theory was to be constructed. It is no accident that the formal introduction of the concept into economics is associated with those very writers whose names are closely connected with the foundation of 'economic science'. It could even be argued that its introduction marks the foundation of the discipline itself, since its appearance divides quite neatly the subsequent literature from the many analyses of individual problems which dominated prior to Smith and the Physiocrats.

Cementing this tradition, Ricardo spoke of fixing his 'whole attention on the permanent state of things' which follows from given changes, excluding for the purposes of general analysis 'accidental and temporary deviations' (1817, p. 88). Marshall, though substituting the terminology 'long-run normal conditions' for the older 'natural conditions', excluded from this category results upon which 'accidents of the moment exert a preponderating influence' (1890, p. vii). J.B. Clark followed suit and held that 'natural or normal' values are those to which 'in the long run, market values tend to conform' (1899, p. 16). Jevons (1871, p. 86), Walras (1874–7, p. 380), Böhm-Bawerk (1899, vol. 2, p. 380) and Wicksell (1901, vol. 1, p. 97) all followed the same procedure.

Not only was the status of 'equilibrium' as the centre of gravitation of the system (the benchmark case, so to speak) preserved, but it was defined in the manner of Smith. The primary theoretical object of all these writers was to explain that situation characterized by a uniform rate of profit on the supply price of capital invested in different lines of production. Walras, whose argument is quite typical, stated the nature of the connection forcefully:

> uniformity of . . . the price of net income [rate of profit] on the capital goods market . . . [is one] condition by which the universe of economic interests is governed. (1874–7, p. 305)

From an historical point of view, the novelty of these arguments which were worked out in the 18th century by Smith and the Physiocrats is not that they recognized that there might be situations which could be described as 'natural', but that they associated these conditions with the outcome of a specific process common to market economies (free competition) and utilized them in the construction of a general economic analysis of market society. Earlier applications of 'natural order' arguments were little more than normative pronouncements about some existing or possible state of society. They certainly made no 'scientific' use of the idea of systematic tendencies, even if these might have been involved. This is particularly apparent in the case of the 'natural law' philosophers, but is also true of the early liberals like Locke and Hobbes. Even Hume, who to all intents and purposes had in his possession all of the building blocks of Smith's position, drew back from the one crucial step that would have led him to Smith's 'method' – he was just not prepared to admit that thinking in terms of regularities, however useful it might prove to be in dispelling theological and other obfuscations (and thus in advancing 'human understanding'), was anything more than a convenient and satisfying way of thinking. The question as to whether the social and economic world was actually governed by such regularities, so central to Smith and the Physiocrats, just did not concern Hume.

Yet the earlier normative connotations of ideas like 'natural conditions', 'natural order', and the like, quite rapidly disappeared when the terminology was appropriated by economic theory. Nothing was 'good' simply by virtue of its being 'natural'. This, of course, is not to say that once the theoretical analysis of the natural tendencies operating in market economies had been completed, and the outcomes of the competitive process had been isolated in abstract, an individual theorist might not at that stage wish to draw some conclusions about the 'desirability' of its results (a normative statement, so to speak). But such statements are not implied by the concept of equilibrium – they are value judgements about the characteristics of its outcomes.

Indeed, contrary to the view sometimes expressed, even Smith's use of deistic analogies and metaphors in the *Theory of Moral Sentiments*, where we read about God as the creator of the 'great machine of the universe', and where we encounter for the first time the famous 'invisible

hand', is no more than the extraneous window-dressing which surrounds a well-defined *theoretical* argument based upon the operation of the so-called 'sympathy' mechanism. Thus, as W.E. Johnson noted when writing for the original edition of Palgrave's *Dictionary*, 'the confusion between scientific law and ethical law no longer prevails', and he observes that 'the term normal has replaced the older word natural' – to be understood by this terminology as 'something which presents a certain empirical uniformity or regularity' (Palgrave 1899, p. 139).

While 'natural conditions' or 'long-run normal conditions' represent the original concept of 'equilibrium' utilized in economic theory, John Stuart Mill's *Political Economy* seems to have been the source from which the actual term equilibrium gained widespread currency (though, like so much else, it is also to be found in Cournot's *Recherches*). More significant, however, is the fact that in Mill's hands the meaning and status of the concept undergoes a modification. While maintaining the idea of equilibrium as a long-period position, Mill introduces the idea that the equilibrium theory is essentially 'static'. The relevant remarks appear at the beginning of the fourth book:

> We have to consider the economical condition of mankind as liable to change … thereby adding a theory of motion to our theory of equilibrium – the Dynamics of political economy to the Statics. (Mill 1848, IV.i, p. 421)

Since he retained the basic category of 'natural and normal conditions', Mill's claim had the effect of adding a property to the list of those associated with the concept of equilibrium. However, over the question of whether this additional property was necessary to the concept of equilibrium, there was to be less uniformity of opinion. Indeed, this matter gave rise to a debate in which at one time or another (until at least the 1930s) almost all theorists of any repute became contributors. The problem was a simple one – are natural or long-period normal conditions the same thing as the 'famous fiction' of the stationary or steady state. Much hinged upon the answer; a 'yes' would have limited the application of equilibrium to an imaginary stationary society in which no one conducts the daily business of life.

On this question, as might be expected, Marshall vacillated. The thrust of his argument (as well as those of his major contemporaries, with the important exception of Pareto) seems to imply that such a property was not essential to his purpose, but as was his habit on so many occasions, in a footnote he qualified that position (1890, p. 379, n.1). In the final analysis, the answer seems to have depended rather more on the explanation given for the determination of equilibrium values, than upon the concept of equilibrium proper. It was not until the 1930s that the issue seems to have been resolved to the general satisfaction of the profession. But then its 'resolution' required the introduction of a new definition of equilibrium (the concept of intertemporal equilibrium) due in the main to Hicks.

However, some further embellishments and modifications were worked upon the concept of equilibrium before the 1930s. Here, two developments stand out. The first concerns the distinction between partial equilibrium analysis and general equilibrium analysis. The second concerns a trend that seems to have developed consequent upon Marshall's treatment of the element of time, which led him to his threefold typology of periods ('market', 'short', and 'long' – we shall leave to one side the further category of 'secular movement'). The upshot of this trend which is decisive, is that it became common to speak of the possibility of 'equilibrium' in each of these Marshallian periods.

The analytical basis for partial equilibrium analysis was laid down in 1838 by Cournot in his *Recherches.* Mathematical convenience, more than methodological principle, seems to have been responsible for his adopting it (see, for example, 1838, p. 127). Though this small volume failed to exercise any widespread influence on the discipline much before the 20th century, it was known and read by Marshall (who spoke of Cournot as his 'gymnastics master'), from whose *Principles* the popularity of partial equilibrium analysis is largely derived (though it would be remiss to overlook Auspitz, Lieben and

von Mangoldt). Unlike the case of Cournot, however, it would be difficult to argue that Marshall came across the method in anything other than a roundabout way (though some have argued that its principal attraction for him lay in its facility in allowing him to express his theory in a manner which required little recourse to mathematics).

When Marshall first introduced the idea of assuming 'other things equal' in the *Principles*, the *ceteris paribus* condition which is taken as the hallmark of the partial equilibrium approach, he seems to have done so not in order to justify the procedure of analysing 'one bit at a time', but in order to make a quite different point – that a long-run normal equilibrium would only *actually* emerge if none but the most general causes were allowed to operate without interference (see, for example, 1890, pp. 36, 366, and 369–70). In other words, the 'other things' that were being held 'equal' were the given data of the theory and the external environment – if the data remained the same and the external environment was freely competitive, then a long-run normal equilibrium would result. Indeed, Walrasian general equilibrium holds 'other things equal' in this sense. To put it another way, in Marshall's initial argument nothing was said about the possibility of assuming the interdependencies between long-run variables themselves to be of secondary importance, as is customary in partial equilibrium analysis.

This latter requirement of Marshallian analysis, the idea of the negligibility of indirect effects when one looks at individual markets (1919, p. 677ff.), seems to have sprung from his habit of presenting equilibrium *theory* in terms of *particular* market demand and supply curves (with their attendant notions of representative consumers and firms). It is here, in fact, that Marshall's presentation of demand and supply theory differs so markedly from its presentation by Walras. To the extent that this is so, it would seem to be better to recognize that the idea of 'partial' versus 'general' equilibrium has more to do with the presentation of a particular theory, and Marshall's propensity to consider markets one at a time, than it has to do with the abstract category of equilibrium with which this discussion is concerned. This view would accord, incidentally, with the fact that the great disputes over the relative merits of these two modes of analysis (for example, that between Walras on the one hand, and Auspitz and Lieben on the other) were fought over the specification of demand and cost functions.

Another modification to the concept of equilibrium that has become more significant in recent literature also makes an appearance in Marshall; though it is not carried as far as it has been in recent literature. The second, third and fifth chapters of the fifth book of Marshall's *Principles* set out the conditions for the determination of what he calls the 'temporary equilibrium', the 'short-run equilibrium' and the 'long-run equilibrium' of demand and supply. The last of these categories, as Marshall makes perfectly clear in the text, corresponds to Adam Smith's 'natural conditions' (1890, p. 347). The first two are to a greater or lesser degree 'more influenced by passing events, and by causes whose action is fitful and short lived' (p. 349). What is striking about Marshall's terminology is the fact that situations which from an analytical point of view would traditionally have been regarded as 'deviations' from long-period normal equilibrium (that is, disequilibria) are explicitly referred to as different cases of 'equilibrium'. This trend has taken on an entirely new significance in recent literature, and has had dramatic consequences for the meaning and status of the concept of equilibrium in economic theory. But just as important in comprehending this development is the introduction of the notion of intertemporal equilibrium into theoretical discourse.

The notion of intertemporal equilibrium (introduced by Hayek, Lindahl and Hicks in the inter-war years and developed in the 1950s by Malinvaud, Arrow and Debreu) warrants special consideration since 'equilibrium conditions' under this notion are defined quite differently from 'natural' or 'long-run normal' conditions. Intertemporal equilibrium defines as its object the determination of nt market-clearing prices (for n commodities over t elementary time periods commencing from an arbitrary short-period

starting point). The chief implication of this definition of equilibrium conditions, and that which sets it apart from long-run normal conditions, is that not only will the price of the same commodity be different at different times but also that the stock of capital need not yield a uniform return on its supply price.

This fundamental change in the concept of equilibrium did not mean that intertemporal equilibrium positions were immediately divested of the status that had been given to 'equilibrium' ever since Adam Smith. In certain circles they continued to be regarded as positions towards which the economic system could actually be said to be 'tending' (or as benchmark cases).

However, once the *sequential* character of this equilibrium concept came to be better understood, it became apparent that there could be no 'tendency' towards it – at least not in the former meaning of that idea. One was either in it, in which case the sequence was 'inessential', or one was not, in which case the sequence was 'essential' (see Hahn 1973, p. 16). And the probabilities overwhelming suggested the latter. Attention was thus turned to the individual points in the sequence; the temporary equilibria, as Hicks had dubbed them (applying the terminology of Marshall in a new context). A whole new class of cases, disequilibrium cases from the point of view of full intertemporal equilibrium, began to be examined. The discipline has now accumulated so many varieties that it is impossible to document them all here. Instead, two broad features of this development may be noted here, the first concerning the role that expectations were thereby enabled to play, the second the common designation now uniformly applied to all such cases: 'equilibrium'.

When equilibrium is interpreted as a solution concept in the sense that *all* solutions to *all* models (for which solutions exist) enjoy equal analytical status and differ only in that they become 'significant', as von Neumann and Morgenstern put it, when they are 'similar to reality in those respects which are essential in the investigation at hand' (1944, p. 32), it is sometimes said that economics has availed itself of a very powerful notion of equilibrium. On this line of argument, Walrasian equilibrium and, say, conjectural equilibrium compete with one another not for the title 'general' (since, in the traditional sense at least, there is no such category), but for the title 'significant'. Furthermore, at any given time they are competing for this title with as many other models as are available to the profession.

It seems to be the case that the status of equilibrium in economic analysis has come full circle since its introduction in the late 18th century. From being derived from the idea that market societies were governed by certain systematic forces, more or less regular in their operation in different places and at different times, it now seems to be based on an opinion that nothing essential is 'hidden' behind the many and varied situations in which market economies might actually find themselves. In fact, it seems that these many cases are to be thought of as being more or less singular from the point of view of modern theory. From being the central organizing category around which the whole of economic theory was constructed, and therefore the ultimate basis upon which its practical application was premissed, equilibrium has become a category with no meaning independent of the exact specification of the initial conditions for *any* model. Instead of being thought of as furnishing a theory applicable, as Mill would have said, to the whole class of cases under consideration, it is increasingly being regarded by theorists as the solution concept relevant to a particular model, applicable to a limited number of cases. The present fashion for replacing economic theory proper by game theory, an approach which could be regarded by no less a theorist than Professor Arrow as contributing only 'mathematical tools' to economic analysis not many years ago (1968, p. 113), seems to exemplify the trend of modern economics.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Competition, Classical
- ▶ Conjectural Equilibria
- ▶ General Equilibrium
- ▶ Temporary Equilibrium

Bibliography

Arrow, K.J. 1968. Economic equilibrium. In *International encyclopedia of the social sciences*, as reprinted in *The collected papers of Kenneth J. Arrow*, vol. 2. - Cambridge, MA: Harvard University Press.

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

Cournot, A.A. 1838. *Researches into the mathematical principles of the theory of wealth.* Trans. N.T. Bacon with an introduction by I. Fisher, 1897; 2nd ed. London/New York: Macmillan. 1927.

Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value. In *Modern capital theory*, ed. M. Brown et al. Amsterdam: North-Holland.

Hahn, F.H. 1973. *On the notion of equilibrium in economics*. Cambridge: Cambridge University Press.

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

Jevons, W.S. 1871. *Theory of political economy.* Edited from the 2nd edition (1879) by R.D.C. Black. Harmondsworth: Penguin, 1970.

Marshall, A. 1890. *Principles of economics*, 9th (variorum) edition, taken from the text of the 8th edition, 1920. London: Macmillan.

Marshall, A. 1919. *Industry and trade*, 2nd ed. London: Macmillan.

Mill, J.S. 1844. *Essays on some unsettled questions of political economy*, 2nd ed. 1874; Reprinted. New York: Augustus M. Kelley.

Mill, J.S. 1848. *Principles of political economy.* 1871 (Reprinted 1909), 6th ed. London: Longmans, Green & Company.

Palgrave, R.H.I., ed. 1899. *Dictionary of political economy*, vol. 3. London: Macmillan.

Pareto, V. 1909. *Manual of political economy.* Trans. from the French edition of 1927 and ed. A.S. Schwier and A.N. Page. New York: Augustus M. Kelley, 1971.

Ricardo, D. 1817. *The principles of political economy and taxation.* Edited from the 3rd edition of 1821 by P. Sraffa with the collaboration of M. Dobb, vol. 1 of *The works and correspondence of David Ricardo*, 11 vols. Cambridge: Cambridge University Press, 1951–73.

Smith, A. 1759. *The theory of moral sentiments.* Edited by D.D. Raphael and A.L. Macfie from the 6th edition of 1790. Oxford: Oxford University Press, 1976.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 2 vols., ed. E. Cannan. London: Methuen, 1961.

Smith, A. 1795. *Essays on philosophical subjects*. Edited by W.P.D. Wrightman and J.C. Bryce. Oxford: Oxford University Press, 1980.

von Böhm-Bawerk, E. 1899. *Capital and interest*, 3 vols. Reprinted. South Holland: Libertarian Press. 1959.

von Neumann, J, and Morgenstern, O. 1944. *Theory of games and economic behaviour*, 3rd ed. Princeton: Princeton University Press, 1953.

Walras, L. 1874–7. *Elements of pure economics*. Trans. and ed. W. Jaffé from the definitive edition of 1926. London: Allen & Unwin, 1954.

Wicksell, K. 1901. *Lectures on political economy*, 2 vols., ed. L. Robbins. London: Routledge and Kegan Paul, 1934.

Equilibrium: An Expectational Concept

Edmund S. Phelps

Economic equilibrium, at least as the term has traditionally been used, has always implied an outcome, typically from the application of some inputs, that conforms to the expectations of the participants in the economy. Many theorists, especially those employing the 'economic man' postulate, have also required the further condition for equilibrium that every participant be optimizing in relation to those correct expectations. However it is the former condition, correct expectations, that appears to be the essential property of equilibrium at least in the orthodox use of the term. Economic equilibrium is therefore not defined in the same terms as physical equilibrium. The rest positions or damped oscillations of pendulums cannot be economic equilibria nor disequilibria since pendulums have no expectations.

Yet it is natural and obvious that the first applications of the equilibrium idea identified some position of rest, or stationary state, as being the equilibrium in the problem at hand. Undoubtedly the term equilibrium, referring to an 'equal weight' of forces pushing capital or what-not *in* as pulling it *out*, owes its origins to the balance of forces prevailing in a stationary situation. But there can also be a *sequence* of positions in which there is a new balance with each new position. There was no reason why equilibria might exist only among stationary states or balanced-growth paths.

Once efforts began to extend economic theory to the case of moving equilibrium paths the

expectational meaning of equilibrium began to be explicit. Two of the pioneers here are Myrdal and Hayek. In his 1927 book on price determination and anticipations (in Swedish) Myrdal addresses the two-way interdependency arising in a dynamic analysis of an on-going economy: present disturbances influence future prices and anticipations of future disturbances affect present prices (the latter relation being Myrdal's main subject). In a 1928 article (in German) on what he called intertemporal equilibrium, Hayek drew the analogy between intertemporal trade and international (or interspatial) trade: prices of the same thing at two different places or times are not generally equal, though they may be pulled up or down together. In a 1929 article (in Swedish) Lindahl studied what is considered to be the first mathematical model of intertemporal equilibrium. This literature is surveyed in Milgate (1979).

The English-speaking world was slow to take up the new line of research. In his *General Theory* of 1936, Keynes speaks grandly of having shown the existence of an (implicitly moving) equilibrium with underemployment, and he does argue that the expectation of falling wages and thus prices makes the slump worse, which suggests he had an expectational notion of equilibrium in mind; but he gives no clues as to what he means by equilibrium, so both the nature and the basis of his claim are left unclear. The new topic of intertemporal equilibrium and the explicit expectational treatment of equilibrium make their English debut in Hicks's *Value and Capital* in 1939. (In the same year Harrod's expectational notion of 'warranted growth', alias equilibrium, and the translation of Lindahl's writings appear.) Hicks makes clear the analytical problem that the analyst and the economic agents alike must solve to find equilibrium: in view of the dependence of future endogenous variables, such as next period's price, on present actions of firms and households, and the dependence of such actions on expectations of those future variables, what expectation would cause the actual outcome to coincide with the expectation? For example, if the actual price P is a function f of the expected price P^e find the value of P^e such that $P^e = f(P^e)$. Thus the fixed-point character of equilibrium from a mathematical standpoint has a human, or real, interpretation. One might say, semi-jocularly, that pendulums have no economic equilibria since their motions, unlike those of trapeze artists, are not a function of expectations, if they have any.

In the postwar period the notion of equilibrium turns up in contexts quite different from that of the inter-war economic theorists. In game theory, begun by von Neumann and Morgenstern, the term equilibrium is used to refer to the theoretical solution to the policies, or play, of two or more players in strategic interaction. If the model postulates optimizing, or expected-utility-maximizing, behaviour by all players, as game theorists' models invariably do, the equilibrium necessarily has the feature that no player can do better acting alone; but lying behind this feature is the essential property that each player has correctly expected the strategy of the others and hence optimized relative to those correct expectations.

In the late 1960s the notion of equilibrium begins to take root in the new territory of non-classical markets – markets without costless and thus complete information. An economy may have markets – the resort hotel market is perhaps a suitable example – in which there are costs in the acquisition or processing of information about prices (and perhaps product specifications) so that arbitrage tendencies are delayed and the classical law of one price operates only with a lag. One well-known portrait of such a market imagines that the national market is composed of Phelpsian islands lacking current-period information about one another's prices. Another image visualizes each firm as an island unto itself with its own stock of customers, who are not knowledgeable about the policies (and perhaps even the whereabouts or existence) of other firms. In such non-Walrasian markets the prevailing prices can be (and usually are) supposed to be market-clearing: no buyer or seller is subjected to rationing (sometimes called non-price rationing by overfastidious writers). However the market will be in *equilibrium* if and only if the prices (and other variables) reflect correct expectations on the part of suppliers and buyers about the prices

prevailing elsewhere – at other islands or other firms; otherwise there is *dis*equilibrium.

An economy may also have markets – one may think of labour markets or markets for rental housing – in which, although information is immediate, the wage or rental setters have to make decisions of some durability, however short-lived, and without advance information about the similar decisions of the other firms. In such quasi-Walrasian markets there may be reasons – having to do with incentives, or efficiency – why wages tend to exceed and rentals lie below the market-clearing level. Yet the market will be in *equilibrium* in the case (if such exists) in which no wage setter or rental setter experiences surprise at the corresponding decisions being made simultaneously (or perhaps somewhat later within the period of the commitment) by the other wage or rental setters; otherwise the market must be in disequilibrium, however long or brief (see Phelps et al. 1970).

Thus the analogy between intertemporal equilibrium and interspatial equilibrium, which was drawn by Hayek and others in their analysis of the former, now seems deeper than it could have at first. The expectational meaning of equilibrium, which is so unavoidably clear in the context of intertemporal equilibrium, where future prices are generally expected future prices, turns out to be just as natural and inevitable in the interspatial context as soon as one gives up the fictive device of the Walrasian auctioneer and thus admit that there are 'other' prices elsewhere, about which there must be expectations, not merely a single market-wide price.

The 1970s witnessed the formal analysis of equilibrium in terms of expectations, or forecasts, of the probability distributions of prices. Lucas, adopting the device of separate market-clearing islands, analysed a model in which there is non-public, or local, information (later called asymmetric information), namely local prices, and these price observations are used to update people's conditional forecasts of the currently unobserved prices elsewhere. There may exist a *rational-expectations* equilibrium in which everyone knows and uses the correct *conditional* expectations of the unobserved prices – that is, the statistically optimal forecasts conditional upon his particular information set. This is equilibrium with a qualification.

In surveying the meaning of equilibrium Grossman has remarked that, in Hicks, 'perfect foresight is an equilibrium concept rather than a condition of individual rationality'. A similar comment applies, with even greater weight, to statistical equilibrium and to its rational-expectations variant. The agents of equilibrium models are not simply rational creatures; they have somehow come to possess fantastic knowledge. The equilibrium premise raises obvious problems of knowledge: why should it be supposed that all the agents have hit upon the true model, and how did they manage to estimate it and conform to it more and more closely? There has always been a strand of thought, running from Morgenstern in the 1930s to Frydman in the present, that holds that we cannot hope to understand the major events in the life of an economy, and perhaps also its everyday behaviour, without entertaining hypotheses of disequilibrium.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Conjectural Equilibria
- ▶ Disequilibrium Analysis
- ▶ General Equilibrium
- ▶ Sunspot Equilibrium
- ▶ Uncertainty and General Equilibrium

Bibliography

Frydman, R., and E.S. Phelps (eds.). 1983. *Individual forecasts and aggregate outcomes*. Cambridge: Cambridge University Press.

Grossman, S.J. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 54: 541–560.

Hayek, F.A. 1928. Das intertemporale Gleichgewichtssystem der Preise und die Bewegungen des Geldwertes. *Weltwirtschaftliches Archiv* 28(1): 33–76.

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

Keynes, J.M. 1936. *General theory of employment, interest and money*. London: Macmillan.

Lindahl, E. 1929. Prisbildningproblemets uppläggning från kapitalteoretisk synpunkt. *Ekonomisk Tidskrift* 2.
Lucas Jr., R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2): 103–124.
Milgate, M. 1979. On the origin of the notion of 'intertemporal equilibrium'. *Economica* 46(1): 1–10.
Morgenstern, O. 1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* 6(3): 337–357.
Myrdal, G. 1927. *Prisbildningsproblemet och Föränderligheten*. Uppsala: Almqvist and Wiksell.
Phelps, E.S., et al. 1970. *Microeconomic foundations of employment and inflation theory*. New York: W.W. Norton.
von Neumann, J., and O. Morgenstern. 1944. *The theory of games*. Princeton: Princeton University Press.

Equilibrium-Correction Models

David F. Hendry

Abstract

The equilibrium-correction class of econometric models is surprisingly large, and includes regression equations, autoregressive-error models, autoregressive distributed-lags, simultaneous equations, autoregressive conditional heteroskedastic processes and generalized ARCH, vector autoregressions and dynamic stochastic general equilibrium systems, among others. Moreover, its properties are relatively generic for all members. Following an historical overview of its origins in error corrections and control mechanisms on the one hand and cointegration on the other, its properties are described, leading to an explanation as to why the ubiquitous class of equilibrium-correction models is prone to forecast failure in processes that are non-stationary from location shifts.

Keywords

Adjustment costs; Autoregressive distributed-lag models; Autoregressive-error models; Cointegration; Common factors; Control mechanisms; Differencing; Dynamic stochastic general equilibrium (DGSE) models; Equilibrium-correction models; Error-correction models; Forecast failure; GARCH processes; Linear-quadratic models; Partial equilibrium; Stationarity; Unit roots; Vector autoregressions

JEL Classifications

C32

Introduction

An equilibrium is a state from which there is no inherent tendency to change. Since we deal with stochastic processes, the equilibrium is the expected value of the variable in an appropriate representation, since that is the state to which the process would revert in the absence of further shocks. Then, we define an equilibrium-correction model (EqCM) as one (*a*) which has a well-defined equilibrium, and (*b*) in which adjustment takes place towards that equilibrium. A key aspect of an EqCM is that deviations from its expected value are attenuated, and eventually eliminated if no additional outside influences impinge. As such, equilibrium-correction models are a very broad class, comprising all regressions, autoregressions, autoregressive-distributed lag (ADL) models, linear simultaneous equations, vector autoregressions (VARs), vector equilibrium-correction systems based on cointegration (VEqCMs), dynamic stochastic general equilibrium systems (DSGEs), autoregressive conditional heteroscedastic processes as in Engle (1982) (ARCH), and generalized ARCH (GARCH, see Bollerslev 1986) processes among others. Their formulation (in levels or differences) determines the equilibrium to which they converge (level or steady state). For example, a random walk without drift is a non-stationary process in levels, but is stationary in differences (its non-integrated representation), and has an expectation of zero, so the differences equilibrium corrects to zero.

We first address the broad nature of the equilibrium-correction class in section "The

Equilibrium-Correction Class", then review the history of equilibrium-correction model formulation in section "Historical Overview", and consider its links to cointegration in section "Equilibrium-Correction and Cointegration". The roles of cointegration and equilibrium correction in economic forecasting are examined in section "Equilibrium Correction and Forecast Failure", in particular the non-robustness of EqCMs to location shifts in the underlying equilibria, and consequently their proneness to forecast failure. Section "Conclusion" concludes.

The Equilibrium-Correction Class

Often it is not realized that the model being used is a member of the equilibrium-correction class, so this section establishes that the models listed above are indeed in the EqCM class. The properties of the class are partly specific to the precise model, but primarily generic, as section "Equilibrium Correction and Forecast Failure" emphasizes. We consider six cases.

Regression as an Equilibrium-Correction Model

Consider a conditional linear equation of the form in (1) for $t = 1; \ldots ; T$:

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i z_{i,t} + \varepsilon_t = \beta_0 + \beta' \mathbf{z}_t + \varepsilon_t \quad (1)$$

with $\varepsilon_t \sim \mathrm{IN}\left[0, \sigma_\varepsilon^2\right]$ (normally and independently distributed, mean zero, variance σ_ε^2) independently of the past and present of the k regressors $\{\mathrm{z}_t\}$. Then:

$$\mathrm{E}[(y_t - \beta_0 - \beta' \mathbf{z}_t) | \mathbf{z}_t] = 0 \quad (2)$$

defines the conditional equilibrium, where adjustment to that equilibrium is instantaneous as entailed by (1). Re-expressing (1) in differences ($\Delta x_t = x_t - x_{t\text{-}1}$ for any x) and lagged deviations from (2) delivers the (isomorphic) EqCM formulation:

$$\Delta y_t = \beta' \Delta \mathbf{z}_t - (y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1}) + \varepsilon_t \quad (3)$$

where the feedback coefficient is -1. Then (3) is an EqCM where the equilibrium-correction term is ($y_{t\text{-}1}$ - β_0 - $\beta' \mathbf{z}_{t\text{-}1}$). Notice that differencing is a linear transformation, not an operator, in any setting beyond a scalar time series.

The existence of (2) does not require that y_t and $\mathbf{z}_t$ are stationary, provided the linear combination is; and could hold, for example, for growth rates rather than the original levels if y_t and $\mathbf{z}_t$ were differences of those original variables.

Autoregressive-Error Models as Equilibrium-Corrections

Even extending a static regression like (1) by (say) a first-order autoregressive error as in:

$$y_t = \beta_0 + \beta' \mathbf{z}_t + u_t \text{ where } u_t = \rho u_{t-1} + \varepsilon_t \quad (4)$$

leads to:

$$y_t = \beta_0 + \beta' \mathbf{z}_t + \rho(y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1}) + \varepsilon_t$$

or:

$$\Delta y_t = \beta' \Delta \mathbf{z}_t + (\rho - 1)(y_{t-1} - \beta_0 - \beta' \mathbf{z}_{t-1}) + \varepsilon_t \quad (5)$$

showing that the common-factor model class (see Sargan 1980; Hendry and Mizon 1978) is also a restricted equilibrium-correction mechanism, constrained by the impact effects (from $\Delta \mathbf{z}_t$) being the same as the long-run effects (from $\mathrm{z}_{t\text{-}1}$).

ADLs as Equilibrium-Correction Models

A first-order autoregressive distributed-lag (ADL) model is:

$$y_t = \beta_0 + \beta_1' \mathbf{z}_t + \beta_2 y_{t-1} + \beta_3' \mathbf{z}_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim \mathrm{IN}\left[0, \sigma_\varepsilon^2\right]. \quad (6)$$

The error $\{\varepsilon_t\}$ on (6) is an innovation against the available information, and its serial independence is part of the definition of the model, whereas normality and homoscedasticity are just for convenience. The condition $|\beta_2| < 1$ is needed to

E

ensure a levels' equilibrium solution: Ericsson (2007) provides an extensive discussion. We consider (6) for both stationary and integrated $\{\mathbf{z}_t\}$, the latter denoting that some of the $\mathbf{z}_t$ have unit roots in their levels representations, but are stationary in differences.

First, under stationarity, taking expectations in (6) where $\mathrm{E}[y_t] = y^*$ and $\mathrm{E}[\mathbf{z}_t] = \mathbf{z}^* \; \forall t$.

$$E[(1-\beta_2)y_t - \beta_0 - (\beta_1 + \beta_3)'\mathbf{z}_t] = 0 \tag{7}$$

so:

$$\begin{aligned} y^* &= \frac{\beta_0}{1-\beta_2} + \frac{1}{1-\beta_2}\,(\beta_1 + \beta_3)'\mathbf{z}^* \\ &= \kappa_0 + \kappa_1'\mathbf{z}^*. \end{aligned} \tag{8}$$

Since many economic theories have long-run partial equilibria like (8), they could be modelled by this class. Transforming (6) to differences and the equilibrium-correction term $\left(y - \kappa_0 - \kappa_1'\mathbf{z}\right)_{t-1}$ delivers:

$$\begin{aligned} \Delta y_t = {} & \beta_1'\Delta\mathbf{z}_t \\ & + (\beta_2 - 1)\left(y_{t-1} - \kappa_0 - \kappa_1'\mathbf{z}_{t-1}\right) \\ & + \varepsilon_t. \end{aligned} \tag{9}$$

The immediate impact of a change in z_t on y_t is β_1, and the rapidity with which Δy_t converges to zero, which is its equilibrium outcome under stationarity, depends on the magnitude of $(\beta_2 - 1) < 0$; when both changes and ε_t are zero (their expectations), (7) results.

When y_t and z_t are integrated of order 1 (denoted I(1)), so are stationary in differences, the reformulation in (9) remains valid provided $|\beta_2| < 1$ in which case $\left(y - \kappa_0 - \kappa_1'\mathbf{z}\right)$ is a cointegration relation, as discussed in section "Equilibrium-Correction and Cointegration". Let $\mathrm{E}[\Delta\mathbf{z}_t] = \delta$ (say) so $\mathrm{E}[\Delta y_t] = \kappa_1'\delta = g_y$ where $\mathrm{E}\left[y_t - \kappa_1'\mathbf{z}_t\right] = \mu$, then taking expectations in (9) using (7):

$$g_y = \beta_1'\delta + (\beta_2 - 1)(\mu - \kappa_0) \tag{10}$$

and subtracting (10) from (9) delivers:

$$\begin{aligned} \Delta y_t = {} & g_y + \beta_1'(\Delta\mathbf{z}_t - \delta) \\ & + (\beta_2 - 1)\left(y_{t-1} - \kappa_1'\mathbf{z}_{t-1} - \mu\right) \\ & + \varepsilon_t. \end{aligned} \tag{11}$$

Re-specifying deterministic terms as in (11) plays an important role in EqCMs, both by helping to orthogonalize the regressors, and because of the pernicious effects of shifts in μ, a topic addressed in section "Equilibrium Correction and Forecast Failure". It is so well known that the standard error of the mean of an IID random variable is the standard deviation of the data divided by the square root of the sample size that it hardly bears reiterating: except that it is somehow almost always ignored in this context. The standard error of the intercept in an EqCM equation like (11) should, therefore, be $\hat{\sigma}_\varepsilon/\sqrt{T}$ but is often a hundred times larger in reported empirical models, revealing a highly collinear specification (a similar comment applies to VARs). Moreover, a check on the model formulation follows from using sample means to estimate δ and μ, then checking that g_y has a sensible value, which may be given by theory (for example, no autonomous inflation, so $g_y = 0$).

Finally, if $\beta_2 = 1$, (9) equilibrium corrects in differences. An autoregression is the special case where $\beta_1 = \beta_3 = 0$, so is also an EqCM; and partial adjustment is another special case where now $\beta_3 = 0$.

GARCH as an Equilibrium-Correction Model

As a fourth example, consider a non-integrated GARCH(1,1) process for ε_t, where $\mathrm{E}\left[\varepsilon_t^2 \,|\, I_{t-1}\right] = \sigma_t^2$ when I_{t-1} denotes past information, and:

$$\sigma_t^2 = \omega + \alpha_{t+1}^2 + \theta\sigma_{t-1}^2 \tag{12}$$

with $0 < \alpha < 1$, $0 < \theta < 1$ and $0 < \alpha + \theta < 1$. Let $\sigma_t^2 = \varepsilon_t^2 - v_t$, where $\mathrm{E}[v_t] = 0$, then:

$$\varepsilon_t^2 = \omega + (\alpha + \theta)\varepsilon_{t-1}^2 + v_t - \theta v_{t-1} \tag{13}$$

where the equilibrium is:

$$\sigma_\varepsilon^2 \equiv \mathrm{E}\left[\varepsilon_t^2\right] = \frac{\omega}{1-(\alpha+\theta)}. \tag{14}$$

Substituting $\omega = (1-(\alpha+\theta))\sigma_\varepsilon^2$ from (14) into the equation for σ_t^2:

$$\Delta\sigma_t^2 = (\theta-1)\left(\sigma_{t+1}^2 - \sigma_\varepsilon^2\right) + \alpha\left(\varepsilon_{t+1}^2 - \sigma_\varepsilon^2\right). \tag{15}$$

Thus, the change in the conditional variance σ^2 responds less than proportionally ($\theta < 1$) to the previous disequilibrium between the conditional variance and the long-run variance, perturbed by the zero-mean discrepancy between the previous squared disturbance ε_{t-1}^2 and the long-run variance σ_ε^2, so the model equilibrium corrects to σ_ε^2, consistent with (14). ARCH is simply a special case.

VARs as Equilibrium-Correction Models

The fifth example is an n-dimensional VAR with m lags and an innovation error $\varepsilon_t \sim \mathrm{IN}_n[0, \Omega_\varepsilon]$:

$$\mathbf{X}_t = \pi + \sum_{i=1}^{m} \Pi_i x_{t-i} + \varepsilon_t \tag{16}$$

where the nm eigenvalues of the polynomial $|\mathbf{I}_n - \sum_{i=1}^{m} \mathbf{\Pi}_i L^i|$ in L determine the characteristics of the time series. If all the eigenvalues are inside the unit circle, (16) is stationary (when all the parameters are constant and the initial conditions also satisfy the process). In that case, $\mathbf{\Gamma} = \left(I_n - \sum_{i=1}^{m} \mathbf{\Pi}_i\right)$ is invertible and has all its eigenvalues inside the unit circle, so the process equilibrium corrects to $\psi = \mathbf{\Gamma}^{-1}\pi$. To illustrate for $m = 2$, (16) can be expressed as:

$$\Delta\mathbf{x}_t = (\mathbf{\Pi}_1 - \mathbf{I}_n)\Delta\mathbf{x}_{t-1} - \mathbf{\Gamma}(\mathbf{x}_{t-2} - \psi) + \varepsilon_t \tag{17}$$

where $\mathrm{E}[\Delta x_t] = 0$ by stationarity, so $\mathrm{E}[x_t - \psi] = 0$ is indeed the equilibrium to which x_t converges in the absence of further shocks. Conversely, if all the eigenvalues are unity, x_t is I(1) with $\mathbf{\Gamma} = 0$ in (17), so does not equilibrium correct in levels, but does so in the differences (unless their polynomial has further unit roots, making the process doubly integrated, I(2)). Finally, for a combination of eigenvalues inside and on the unit circle, $\mathbf{\Gamma}$ has reduced rank $0 < r < n$ equal to the number of non-unit eigenvalues, so can be expressed as $\mathbf{\Gamma} = \alpha\beta'$ where α and β also have rank r. Then π in (16) can be decomposed into the unconditional growth rate of x_t, denoted γ, and $\alpha\mu$ such that in place of (17), we have:

$$\Delta\mathbf{x}_t = \gamma + (\mathbf{\Pi}_1 - \mathbf{I}_n)(\Delta\mathbf{x}_{t-1} - \gamma) - \alpha(\beta'\mathbf{x}_{t-2} - \mu) + \varepsilon_t \tag{18}$$

so that $\mathrm{E}[\beta' x_t - \mu] = 0$ and the system converges to that equilibrium when the original variables are I(1), hence $\beta' x_{t-2}$ is an I(0) process which equilibrium corrects to μ. At the same time, $\Delta\mathrm{x}_t$ is an I(0) process which equilibrium corrects to γ, noting that $\beta'\gamma = 0$, whereas x_t drifts.

Linear simultaneous equations systems of time series are a restriction on a VAR, so are also EqCMs.

DSGEs as Equilibrium-Correction Models

As a final brief example, well-defined general equilibrium systems have equilibria. Using Taylor-series expansions around the steady-state values of the discretized representation of a system of differential equations, Bårdsen et al. (2004) show that any dynamic system with a steady-state solution has a linear EqCM representation. Thus, they argue that linearizations of DSGEs imply linear EqCM representations. In principle, these could be in terms of changes only, corresponding to a steady-state path. More usually, level solutions result.

Historical Overview

Equilibrium-correction models are a special case of the general class of proportional, derivative and integral control mechanisms, so have a long pedigree in that arena: for economics examples, see Phillips (1954), Phillips and Quenouille (1960) and Whittle (1963), with the links summarized in Salmon (1988). Explicit examples of EqCMs

E

are presented in Sargan (1964) and were popularized by Davidson et al. (1978), although they were called 'error-correction mechanisms' (ECMs) by those authors. The major developments underlying cointegration in Engle and Granger (1987) established its isomorphism with equilibrium correction for integrated processes, leading to an explosion in the application of EqCMs and the development of a formal analysis of vector EqCM systems in Johansen (1988, 1995). We now review the two stages linking control mechanisms with error correction, then that with equilibrium correction.

Error Correction and Control Mechanisms

Phillips (1954, 1957), in particular, pioneered the application of control methods for macroeconomic stabilization, specifically techniques for derivative, proportional and integral control servomechanisms. In this form of control, a target (say an unemployment rate of five per cent) is to be achieved by adjusting an instrument (say government expenditure), and changes to the instrument, its level, and cumulative past errors may need to be included in the rule to stabilize the target.

That approach is a precursor to the well-known linear-quadratic model in which one optimizes a quadratic function of departures from target trajectories for a linear dynamic system over a finite future horizon (see, for example, Holt et al. 1960; Preston and Pagan 1982). For example, consider the quadratic cost function C_H which penalizes the deviations of a variable x_{t+j} from a pre-specified target trajectory $\left\{x^*_{t+j}\right\}$ subject to costs of adjustment from changes $\Delta x_t = x_t - x_t$ over an H-period horizon commencing at time t:

$$C_H = \sum_{j=0}^{H} c_{t+j} = \sum_{j=0}^{H} \frac{1}{2}\left[\left(x_{t+j} - x^*_{t+j}\right)^2 + \alpha\left(\Delta x_{t+j}\right)^2\right]. \tag{19}$$

To minimize c_{t+j} at time $t + j$, differentiate with respect to x_{t+j}, noting the intertemporal link that $\Delta x_{t+j+1} = x_{t+j+1} - x_{t+j}$ also depends on x_{t+j}, which yields (ignoring the end point for simplicity):

$$\frac{\partial C_H}{\partial x_{t+j}} = \frac{\partial c_{t+j}}{\partial x_{t+j}} + \frac{\partial c_{t+j+1}}{\partial x_{t+j}} = x_{t+j} - x^*_{t+j} \alpha\left(\Delta x_{t+j}\right) - \alpha\left(\Delta x_{t+j+1}\right), \tag{20}$$

so equating to zero for a minimum for any j, and hence for $j = 0$:

$$x_t - x^*_t + \alpha \Delta x_t - \alpha \Delta x_{t+1} = 0.$$

Expressed as a polynomial in leads and lags in the operator L (for $\alpha \neq 0$):

$$\left(L^{-1} - \left(2 + \alpha^{-1}\right) + L\right)x_t = \left(L^{-1} - \lambda_2\right)(1 - \lambda_1 L)x_t = -\frac{x^*_t}{\alpha}. \tag{21}$$

The polynomial in (21) has roots λ_1 and λ_2 with a product of unity (so they are inverses, with λ_1 inside and λ_2 outside the unit circle) and a sum of $(2 + \alpha^{-1})$. Inverting the first factor $(L^{-1} - \lambda_2)$, using $(1/\lambda_2) = \lambda_1 < 1$ and expanding the last term as a power series in L^{-1}expresses x_t as a function of lagged x_s and current and future values of x^*_{t+k}:

$$(1 - \lambda_1 L)x_t = \frac{\lambda_1}{\alpha}\left(1 + \lambda_1 L^{-1} + \lambda_1^2 L^{-2} + \cdots\right)x^*_t = \frac{\lambda_1}{\alpha}\sum_{k=0}^{\infty} \lambda_1^k x^*_{t+k}. \tag{22}$$

Since $(1 - \lambda_1) = \lambda_1/\alpha(1 - \lambda_1)$, let:

$$x^{**}_t = (1 - \lambda_1)\sum_{k=0}^{\infty} \lambda_1^k x^*_{t+k} \tag{23}$$

denote the 'ultimate' target (scaled so that the weights sum to unity as in, for example, Nickell 1985) then from (22) using (23), for $t < H$:

$$\Delta x_t = -(1 - \lambda_1)\left(x_{t-1} - x^{**}_t\right) = (1 - \lambda_1)\Delta x^{**}_t - (1 - \lambda_1)\left(x_{t-1} - x^{**}_{t-1}\right). \tag{24}$$

Thus, x_t adjusts to changes in the ultimate target, and to the previous error from that target, and is an EqCM when $-1 < \lambda_1 < 1$. Mistakes in plans, errors in expectations, and relations between the ultimate target and its determinants all need to be modelled for an operational rule. To hit a moving target requires a feedforward rule, and the role of $\alpha(\Delta x_{t+j})^2$ in (19) is to penalize the controller from making huge changes to x_t when doing so. However, it is difficult to imagine real world adjustment costs being proportional to changes, which in any case then depend on the specification of x_t as logs, levels, proportions or even changes (see, for example, Nickell 1985). Moreover, the entire class is partial adjustment, as (24) shows.

For 1-period optimization (so $H = 0$: see, for example, Hendry and Anderson 1977), only the end point is relevant, so (20) delivers the planned value x_t^p as a function of $x_t^{**} = x_t^*$:

$$\begin{aligned} x_t^p - x_{t-1} &= \frac{1}{1-\alpha}\left(x_t^* - x_{t-1}\right) \\ &= \rho\left(x_t^* - x_{t-1}\right). \end{aligned} \tag{25}$$

When the error on the plan is $\varepsilon_t = x_t - x_t^p$, where $\mathrm{E}[x_t^p \varepsilon_t] = 0$ under rationality, and $x_t^* = \beta' \mathbf{z}_t$ (say), (25) becomes:

$$\begin{aligned} \Delta x_t &= \rho(\beta' \mathbf{z}_t - x_{t-1}) + \varepsilon_t \\ &= \rho\beta' \Delta \mathbf{z}_t - \rho(x_{t-1} - \beta' \mathbf{z}_{t-1}) + \varepsilon_t. \end{aligned}$$

This is a partial adjustment again. The static regression in section "Regression as an Equilibrium-Correction Model" has a more restrictive dynamic structure, but otherwise the properties of the ADL in section "ADLs as equilibrium-correction models" can vary over a wide range (see Hendry 1995, ch. 6).

From Error Correction to Equilibrium Correction

The model in Sargan (1964) was explicitly an ECM for wages and prices (w_t and p_t denote their respective logs), building on previous models of wage and price inflation written as:

$$\Delta w_t = \beta_0 + \beta_1 \Delta p_t + \beta_2 \Delta w_{t-1} + \varepsilon_t. \tag{26}$$

When $\mathrm{E}[\varepsilon_t] = 0$ and the differenced variables are stationary with means $\mathrm{E}[\Delta w_t] = \dot{w}$ and $\mathrm{E}[\Delta p_t] = \dot{p}$, then the long-run steady-state solution to (26) is:

$$\dot{w} = \frac{\beta_0 + \beta_1 \dot{p}}{1 - \beta_2}.$$

As formulated, (26) does not establish any relationship between the levels w_t and p_t, hence these could drift apart. Since economic agents are concerned about the level of real wages, $w_t - p_t$, Sargan postulated the equilibrium:

$$(w - p)_{e,t} = \delta_0 + \delta_1 \Delta p_t + \delta_2' \mathbf{z}_t, \tag{27}$$

where $\mathbf{z}_t$ denotes a vector of additional variables, such as unemployment (u), productivity (q) and political factors. The disequilibrium is:

$$v_t = w_t - p_t - \delta_0 - \delta_1 \Delta p_t - \delta_2' \mathbf{z}_t \tag{28}$$

and, to re-establish equilibrium whenever the levels drift apart, he used the explicit adjustment equation:

$$\begin{aligned} \Delta w_t &= \alpha\left(w_{t-1} - p_{t-1} - (w - p)_{e,t-1}\right) \\ &= \alpha v_{t-1}. \end{aligned} \tag{29}$$

If a relation like (28) is well defined with v_t being I(0) when the levels are I(1), so the differences are I(0), then w_t forms a non-integrated combination with p_t and z_t so these variables are cointegrated (see, among many others, Engle and Granger 1987; Phillips and Loretan 1991; Banerjee, et al. 1993.).

A less restricted specification than (26) entails including the levels terms $(w - p)_{t-1}$ and $\mathbf{z}_{t-1}$ (and their differences), so if contemporaneous variables are excluded:

$$\begin{aligned} \Delta w_t = \pi_0 &+ \pi_1 \Delta p_{t-1} + \pi_2 \Delta w_{t-1} \\ &- \pi_3 (w - p)_{t-1} + \pi_4' \mathbf{z}_{t-1} \\ &+ \pi_5' \Delta \mathbf{z}_{t-1} + u_t. \end{aligned} \tag{30}$$

When $\pi_3 \neq 0$, the long-run levels equilibrium solution to (30) matching (27) is ($\phi_4 = \pi_4/\pi_3$):

$$\mathrm{E}\left[w - p - \varphi_4' \mathbf{z}\right\} = f(\dot{w}, \dot{p}, \dot{\mathbf{z}}).$$

The model in (30) has both derivative and proportional control (e.g., Δp_{t-1} and $(w - p)_{t-1}$) following up Phillips (1954, 1957) (see Salmon 1982). The proportional mechanism ensures the disequilibrium adjustment, based on the (possibly detrended) log-ratio of two nominal levels (see, for example, Bergstrom 1962). The equivalent of g_y in section "ADLs as equilibrium-correction models" should be $\dot{p} + \dot{q}$ in (30) to avoid having 'autonomous wage inflation' independent of all economic forces.

The long-run stability of the 'great ratios' in Klein (1953) was often implicitly assumed to justify such transformations, but had come under question (see, for example, Granger and Newbold 1977, and the discussion in Hendry 1977), although Hendry and Mizon (1978) had argued that what mattered was that the errors in (30) were stationary, not that all the variables were stationary. Granger (1981) related the type of model in (30) to cointegration, and Granger (1986) showed the important new result that one of Δw_t or Δp_t must depend on the equilibrium correction if w_t and p_t were cointegrated: the assumption in (29) is that w_t adjusts to the disequilibrium. If both variables w_t and p_t adjust to the disequilibrium, then p_t is not weakly exogenous for the $\{\pi_i\}$ (see Phillips and Loretan 1991; Hendry 1995). It is primarily because of cointegration that equilibrium-correction models like (30) have proved a popular specification. Engle and Granger (1987) showed that cointegration and proportional EqCM were equivalent, linking time-series approaches more closely with econometric modelling. Davidson and Hall (1991) also linked VARs as in section "VARs as Equilibrium-Correction Models" to target relations as discussed in section "Error Correction and Control Mechanisms" using cointegration analysis, so we now turn to the topic of cointegration in more detail.

Equilibrium-Correction and Cointegration

From the ADL to a VAR

To complete (6), a process is needed for $\{\mathrm{z}_t\}$. Let:

$$\mathbf{z}_t \,|y_{t-1}, \mathbf{z}_{t-1} \sim \mathrm{N}_k\left[\pi_{20} + \pi_{21} y_{t-1} + \pi_{22} \mathbf{z}_{t-1}, \Omega_{zz}\right]. \tag{31}$$

Given (31), the joint distribution is the first-order VAR:

$$\begin{pmatrix} y_t \\ \mathbf{z}_t \end{pmatrix} | y_{t-1}, \mathbf{z}_{t-1} \sim \mathrm{N}_{k+1}\left[\begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} + \begin{pmatrix} \pi_{11} & \pi_{12}' \\ \pi_{21} & \pi_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ \mathbf{z}_{t-1} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12}' \\ \sigma_{12} & \Omega_{zz} \end{pmatrix}\right]. \tag{32}$$

Consequently, to match (6):

$$\mathrm{E}[y_t | \mathbf{z}_t, y_{t-1}, \mathbf{z}_{t-1}] = \pi_{10} + \pi_{11} y_{t-1} + \pi_{12} \mathbf{z}_{t-1} + \sigma_{12}' \Omega_{zz}^{-1} (\mathbf{z}_t - \pi_{20} - \pi_{21} y_{t-1} - \pi_{22} \mathbf{z}_{t-1}), \tag{33}$$

so $\beta_0 = (\pi_{10} - \phi' \pi_{20})$, $\beta_1 = \phi$, $\beta_2 = \pi_{11} - \phi' \pi_{21}$ and $\beta_3 = (\pi_{12} - \phi' \pi_{22})$ when $\varphi = \Omega_{zz}^{-1} \sigma_{12}$, and $\sigma_\varepsilon^2 = \sigma_{11} - \sigma'_{12} \Omega_{zz}^{-1} \sigma_{12}$. When $\mathbf{z}_t$ is weakly exogenous for $(\beta_0, \ldots, \beta'_3)$, the model in (31) can be ignored when analysing (6); also $\pi_{21} = 0$ then ensures the strong exogeneity of $\mathbf{z}_t$ for $(\beta_0, \ldots, \beta'_3)$.

Sufficient conditions for stationarity of (32) are that all the eigenvalues λ_i of the matrix of the $\{\pi_{ij}\}$ are inside the unit circle, but a more realistic setting allows for unit roots in π_{22}. On that basis, we now investigate the properties of the VAR in (32) letting $\mathbf{x}_t' = (y_t : \mathbf{z}_t')$ as in (16).

Cointegration

Linear combinations of I(1) processes are usually I(1) as well: differencing is still needed to remove the unit root. Sometimes integration cancels between series to yield an I(0) outcome and

thereby deliver cointegration. Cointegrated processes in turn define a 'long-run equilibrium trajectory' for the economy, departures from which induce 'equilibrium correction' to move the economy back towards its path. A rationale for integrated–cointegrated data is that economic agents use fewer equilibrium corrections than there are variables they need to control. We can see that effect as follows.

Consider the bivariate VAR:

$$\begin{aligned} x_{1,t} &= \pi_{10} + \pi_{11}x_{1,t-1} + \pi_{12}x_{2,t-1} + \varepsilon_{1,t} \\ x_{2,t} &= \pi_{20} + \pi_{21}x_{1,t-1} + \pi_{22}x_{2,t-1} + \varepsilon_{2,t}, \end{aligned} \tag{34}$$

where $(\varepsilon_{1,ts}, \varepsilon_{2,t})$ are bivariate independent normal. To determine when the system is I(1) and if so, whether or not some linear combinations of variables are cointegrated, rewrite (34) as:

$$\begin{aligned} \begin{pmatrix} \Delta x_{1,t} \\ \Delta x_{2,t} \end{pmatrix} &= \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} \\ &+ \begin{pmatrix} (\pi_{11} - 1) & \pi_{12} \\ \pi_{21} & (\pi_{22} - 1) \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} \\ &+ \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \end{aligned} \tag{35}$$

or as (a special case of (18)):

$$\Delta \mathbf{x}_t = \pi + \mathbf{\Pi}\mathbf{x}_{t-1} + \varepsilon_t. \tag{36}$$

Three cases are of interest. First $\pi = 0$, so (36) is a vector random walk without any levels relationships, and so $\mathbf{x}_t$ is I(1) with $\Delta\mathbf{x}_t$ being I(0) and equilibrium correcting to π. Secondly, if $\mathbf{\Pi}$ has full rank, then $\mathbf{x}_t$ is I(0) and equilibrium corrects to $\mathbf{\Pi}^{-1}\pi$. The most interesting case is when Π is reduced rank so can be expressed as:

$$\mathbf{\Pi} = \alpha\beta' = \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \end{pmatrix} (\beta_{11} \quad \beta_{12}),$$

where we will normalize $\beta_{11} = 1$. Then in (35):

$$\begin{aligned} \begin{pmatrix} \Delta x_{1,t} \\ \Delta x_{2,t} \end{pmatrix} &= \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} \\ &+ \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \end{pmatrix} (1 \ \beta_{12}) \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \\ &= \begin{pmatrix} \pi_{10} \\ \pi_{20} \end{pmatrix} + \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \end{pmatrix} (x_{1,t-1} + \beta_{12}x_{2,t-1}) + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \end{aligned} \tag{37}$$

which is an EqCM with $(x_{1,t-1} + \beta_{12}x_{2,t-1})$ stationary. Thus, cointegration entails EqCM and vice versa when the feedback relation is I(0). However, prior to Granger (1981) the EqCM literature did not visualize a single cointegration relation affecting several variables, and thereby making them integrated, but instead just took the non-stationarity of the observed data as due to the behaviour of the non-modelled variables. Consequently, system cointegration 'endogenizes' data integrability in a consistent way, and so represents a significant step forward. The extensive literature on cointegration analysis also addresses most of the estimation and formulation issues that arise when seeking to conduct inference in integrated-cointegrated processes: much of this is summarized in Hendry and Juselius (2001), to which the interested reader is referred for bibliographic perspective.

Equilibrium Correction and Forecast Failure

Recent research on the impact of structural breaks, particularly location shifts, on cointegrated processes has emphasized the need to distinguish equilibrium correction, which operates successfully only within regimes, from error correction, which stabilizes in the face of other non-stationarities (see, for example, Clements and Hendry 1995). The assumptions concerning the stationarity, or otherwise, of the entity to be controlled in section "Error Correction and Control Mechanisms" were rarely explicitly stated, but suggest an implicitly stationary system (or perhaps steady-state growth). In such a setting,

equilibrium-correction or cointegration relationships prevent the levels of the variables from 'drifting apart', and so improve the properties of forecasts.

Practical work, however, must allow the data generation process to be non-stationary both from unit roots (that is, I(1) or possibly I(2)) and from a lack of time invariance. When data processes are non-stationary even after differencing and cointegration, equilibrium-correction mechanisms tend to suffer from forecast failure, defined as a significant deterioration in forecast performance relative to in-sample behaviour. Since most empirical model forms are members of the EqCM class, this is a serious practical problem.

To illustrate, reconsider the special case of (18) with just one lag, written as:

$$\Delta \mathbf{x}_t = \gamma + \alpha(\beta' \mathbf{x}_{t-1} - \mu) + \varepsilon_t. \tag{38}$$

The shift of interest here is $\nabla\mu^* = \mu^* - \mu$, where μ^* denotes the post-break equilibrium mean (reasonable magnitude shifts in γ, α and $\mathbf{\Omega}_\varepsilon$ rarely entail forecast failure). Denote the forecast origin as time T, then following a change to μ^* immediately after forecasting, the next outcome is:

$$\begin{aligned} \Delta \mathbf{x}_{T+1} &= \gamma + \alpha(\beta' \mathbf{x}_T - \mu^*) + \varepsilon_{T+1} \\ &= \gamma + \alpha(\beta' \mathbf{x}_T - \mu) + \varepsilon_{T+1} - \alpha\nabla\mu^* \end{aligned} \tag{39}$$

where $-\alpha\nabla\mu^*$ is the unanticipated break, and becomes the mean forecast error for known parameters. Importantly, the 1-step ahead forecast at $T + 1$ using an unchanged model suffers the same mistake:

$$\begin{aligned} &E[\Delta \mathbf{x}_{T+2} - (\gamma + \alpha(\beta' \mathrm{x}_{T+1} - \mu))] \\ &\quad = -\alpha\nabla\mu^* \end{aligned} \tag{40}$$

so the shift in the equilibrium mean induces systematic mis-forecasting. The impact on multi-step forecasts of the levels is even more dramatic, as the mean forecast error increases at every horizon, eventually converging to $\alpha(\beta'\alpha)^{-1}\nabla\mu^*$, which can be very large (see Clements and Hendry 1999). Thus, EqCMs are a non-robust forecasting device in the face of equilibrium-mean shifts, a comment which therefore applies to all members of this huge class of model, including GARCH (as noted earlier), where the pernicious shift is in the unconditional variance σ_ε^2 in (15).

To avoid forecast failure, more adaptive methods merit consideration. One generic approach to improving robustness to location shifts is to difference the forecasting device (although that may well worsen the impact of large measurement errors at the forecast origin). Differencing can be before estimation, as in a double-differenced VAR, or after, as in differencing the estimated EqCM to eliminate the equilibrium mean and growth intercept. Such devices perform as badly as the EqCM in terms of forecast biases when a break occurs after forecasts are announced (see Clements and Hendry 1999), and have a larger error variance. The key difference is their performance when forecasting after a break has already occurred, in which case the EqCM continues to perform badly (as shown above in (48)), but a DEqCM becomes relatively immune to the earlier break. Taking (47) as an example, an additional difference yields:

$$\begin{aligned} \Delta^2 \mathbf{x}_{T+1} &= \Delta(\gamma - \alpha\mu^*) + \alpha\beta' \Delta \mathrm{x}_T + \Delta\varepsilon_{T+1} \\ &= -\alpha\nabla\mu^* + \alpha\beta' \Delta \mathrm{x}_T + \Delta\varepsilon_{T+1} \end{aligned}$$

so there is no benefit when forecasting immediately after the break (as $\Delta\mu^* = \mathrm{r}\mu^*$), whereas (48) becomes:

$$\Delta^2 \mathbf{x}_{T+2} = \alpha\beta' \Delta \mathbf{x}_{T+1} + \Delta\varepsilon_{T+2}$$

since $\Delta\mu^* = 0$. Thus, there is no longer any systematic failure. The same comment applies to double-differenced devices, although Hendry (2006) shows how to improve these while retaining robustness.

A further consequence is that, when a location shift is not modelled, since most econometric estimators minimize mis-fitting, the coefficients of dynamic models will be driven towards unity, which induces differencing to convert a location shift into a 'blip'. Thus, estimates that apparently manifest 'slow adjustment' may just reflect unmodelled breaks.

An alternative approach to avoiding forecast failure would be to construct a genuine error-correction model, adjusting more or less rapidly to wherever the target variable moves: for example, exponentially weighted moving averages do so for some processes. In essence, either the dynamics must ensure correction or the target implicit in the econometric model must move when the regime alters. This last result also explains why models in differences are not as susceptible to certain forms of structural break as equilibrium-correction systems (again see Clements and Hendry 1999), and in turn helps to account for many of the findings reported in the forecasting competitions literature. When the shift in question is a change in a policy regime, Hendry and Mizon (2005) suggest approaches to merging robust forecasts with policy models.

Conclusion

Equilibrium-correction models have a long pedigree as an 'independent' class, related to optimal control theory. However, their isomorphism with cointegrated relationships has really been the feature that has ensured their considerable popularity in empirical applications. In both cases, part of the benefit from the EqCM specification came from expressing variables in the more orthogonalized forms of differences and equilibrium-correction terms, partly from the resulting insights into both short-run and long-run adjustments, partly from discriminating between the different components of the deterministic terms, and partly from 'balancing' regressors of the same order of integration, namely I(0).

Unfortunately, science is often two steps forward followed by one back, and that backwards step came from an analysis of EqCMs when forecasting in the face of structural breaks. Unmodelled shifts in the equilibrium mean (and less so in the growth rate) induce forecast failure, making EqCMs a non-robust device with which to forecast when data processes are prone to breaks, as many empirical studies suggest they are (see, for example, Stock and Watson 1996). Since cointegration hopefully captures long-run causal relations, and ties together the levels of I(1) variables, eliminating its contribution should not be undertaken lightly, hence the suggestion in section "Equilibrium Correction and Forecast Failure" of using the differenced version of the estimated EqCM for forecasting.

See Also

▶ Cointegration

Acknowledgment Financial support from the ESRC under Professorial Research Fellowship RES051270035 is gratefully acknowledged, as are helpful comments from Gunnar Bårdsen, Julia Campos, Jennifer Castle, Mike Clements, Søren Johansen and Graham Mizon.

Bibliography

Bårdsen, G., S. Hurn, and K.A. Lindsay. 2004. Linearizations and equilibrium correction models. *Studies in Nonlinear Dynamics and Econometrics* 8(4): 5.

Banerjee, A., J.J. Dolado, J.W. Galbraith, and D.F. Hendry. 1993. *Co-integration, error correction and the econometric analysis of non-stationary data*. Oxford: Oxford University Press.

Bergstrom, A.R. 1962. A model of technical progress, the production function and cyclical growth. *Economica* 29: 357–370.

Bollerslev, T. 1986. Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* 51: 307–327.

Clements, M.P., and D.F. Hendry. 1995. Macro-economic forecasting and modeling. *Economic Journal* 105: 1001–1013.

Clements, M.P., and D.F. Hendry. 1999. *Forecasting non-stationary economic time series*. Cambridge, MA: MIT Press.

Davidson, J.E.H., and S. Hall. 1991. Cointegration in recursive systems. *Economic Journal* 101: 239–251.

Davidson, J.E.H., D.F. Hendry, F. Srba, and J.S. Yeo. 1978. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88: 661–692.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1007.

Engle, R.F., and C.W.J. Granger. 1987. Cointegration and error correction: Representation, estimation and testing. *Econometrica* 55: 251–276.

Ericsson, N.R. 2007. *Econometric modeling*. Oxford: Oxford University Press.

E

Granger, C.W.J. 1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16: 121–130.
Granger, C.W.J. 1986. Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics* 48: 213–228.
Granger, C.W.J., and P. Newbold. 1977. The time series approach to econometric model building. In *New methods in business cycle research*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.
Hendry, D.F. 1977. On the time series approach to econometric model building. In *New methods in business cycle research*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.
Hendry, D.F. 1995. *Dynamic econometrics*. Oxford: Oxford University Press.
Hendry, D.F. 2006. Robustifying forecasts from equilibrium-correction models. *Journal of Econometrics* 135: 399–426. Special issue in honor of Clive Granger.
Hendry, D.F., and G.J. Anderson. 1977. Testing dynamic specification in small simultaneous systems: An application to a model of building society behaviour in the United Kingdom. In: *Frontiers in quantitative economics,* vol. 3. ed. M.D. Intriligator. Amsterdam: North-Holland.
Hendry, D.F., and K. Juselius. 2001. Explaining cointegration analysis: Parts I and II. *Energy Journal* 21: 1–42; 22: 75–120.
Hendry, D.F., and G.E. Mizon. 1978. Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England. *Economic Journal* 88: 549–563.
Hendry, D.F., and G.E. Mizon. 2005. Forecasting in the presence of structural breaks and policy regime shifts. In *Identification and inference for econometric models: Essays in Honor of Thomas Rothenberg*, ed. D.W.K. Andrews and J.H. Stock. Cambridge: Cambridge University Press.
Holt, C., F. Modigliani, J.F. Muth, and H. Simon. 1960. *Planning production, inventories and work force*. Englewood Cliffs: Prentice-Hall.
Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12: 231–254.
Johansen, S. 1995. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford: Oxford University Press.
Klein, L.R. 1953. *A textbook of econometrics*. Evanston: Row, Peterson and Company.
Nickell, S.J. 1985. Error correction, partial adjustment and all that: An expository note. *Oxford Bulletin of Economics and Statistics* 47: 119–130.
Phillips, A.W.H. 1954. Stabilization policy in a closed economy. *Economic Journal* 64: 290–333.
Phillips, A.W.H. 1957. Stabilization policy and the time form of lagged response. *Economic Journal* 67: 265–277.
Phillips, P.C.B., and M. Loretan. 1991. Estimating long-run economic equilibria. *Review of Economic Studies* 58: 407–436.
Phillips, A.W.H., and M.H. Quenouille. 1960. Estimation, regulation and prediction in interdependent dynamic systems. *Bulletin de l'Institut de Statistique* 37: 335–343.
Preston, A.J., and A.R. Pagan. 1982. *The theory of economic policy*. Cambridge: Cambridge University Press.
Salmon, M. 1982. Error correction mechanisms. *Economic Journal* 92: 615–629.
Salmon, M. 1988. Error correction models, cointegration and the internal model principle. *Journal of Economic Dynamics and Control* 12: 523–549.
Sargan, J.D. 1964. Wages and prices in the United Kingdom: A study in econometric methodology (with discussion). In: *Econometric analysis for national economic planning*, Colston papers, vol. 16. ed. P.-E. Hart, G. Mills, and J.K. Whitaker. London: Butterworth.
Sargan, J.D. 1980. Some tests of dynamic specification for a single equation. *Econometrica* 48: 879–897.
Sims, C.A., ed. 1977. *New methods in business cycle research*. Minneapolis: Federal Reserve Bank of Minneapolis.
Stock, J.H., and M.W. Watson. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* 14: 11–30.
Whittle, P. 1963. *Prediction and regulation by linear least-square methods*. Princeton: D. Van Nostrand.

Equity

Allan M. Feldman

Depending on the user's inclinations, 'equity' can mean almost anything; this user will adopt a meaning which has been followed by economists and other social scientists since the late 1960s (see particularly Foley 1967), a meaning close to equality or fairness.

Although 'equality' is less ambiguous than 'equity', it too has many definitions: Jefferson's adage that 'all men are created equal' clearly does not mean that they all have the same talents, skills, inherited and acquired wealth; it only means that they share, or ought to share, certain narrowly defined legal rights and political powers. However, in a simple economic model, equality can

be made simple. If we assume that society is comprised of a certain set of n individuals who produce among themselves certain quantities of various goods, we can speak of an equal division of the goods: an allocation that would give each person exactly $1/n$ of the total of each good. Economists would agree that this is equality (at least on the consumption side). Most would also agree that it is an undesirable state of affairs, if for no other reason than that no two people would ever want to consume exactly the same bundle of goods. They would be equal, but not especially happy. Moreover, getting society to that equal allocation would require transferring wealth from the more productive individuals to the less productive, and the transfer mechanism itself would destroy incentives to produce.

So equality in its extreme form – an equal consumption bundle for every consumer – is an obviously unworkable idea, and needs to be weakened. We shall say in this assay that individual i *envies* individual j if i would rather have j's consumption bundle than his own. Formally, let $u_i(\cdot)$ represent individual i's utility function, and x_i represent his consumption bundle. (For now, production is ignored.) Then i envies j if $u_i(x_j) > u_i(x_i)$. This is now a more- or-less standard usage by economists, who have ignored wiser and older counsel, for example, J. S. Mill, who calls envy 'that most odious and anti-social of all passions' (*On Liberty*, ch. 4). Mill would presumably not endorse an economic analysis founded on envy.

Following Varian (1974) we define an allocation as *equitable* if under it no individual envies another; that is, if

$$u_i(x_i) \geq u_i\left(x_j\right) \text{ for all } i \text{ and } j.$$

Obviously, the equal allocation is equitable. But equity does not share equality's obvious disadvantage of forcing all to consume the same no matter what their tastes. If Adam loves ȧpples and Eve loves oranges, and if God has endowed them with a total of one apple and one orange, then the equal allocation (half an apple and half an orange for each) is clearly foolish, but the equitable allocation (one apple for Adam and one orange for Eve) makes good sense.

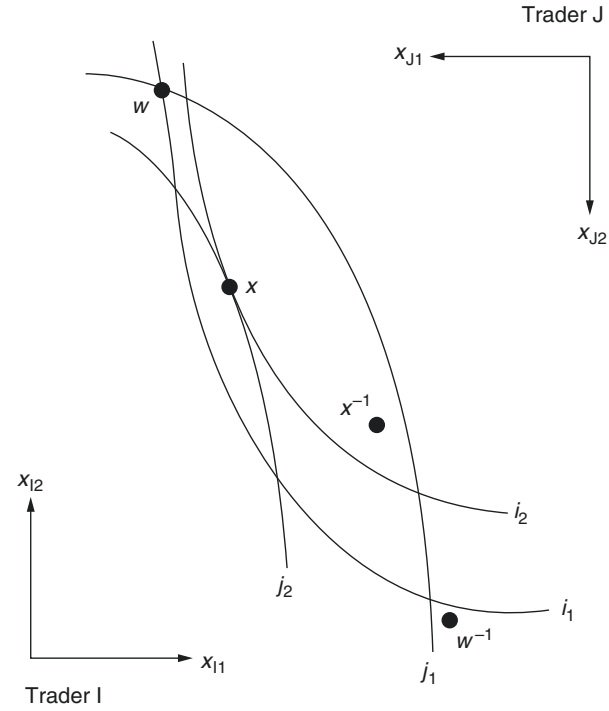
But the notion of equity has an obvious disadvantage, aside from its being founded on that odious passion. For instance, the economist's model, which reduces person i to a utility function $u_i(\cdot)$ and a bundle of goods x_i, ignores the fact that life is full of things not captured in $u_i(\cdot)$ or x_i, for instance, non-transferable attributes like beauty, health and family. Even if the division of economic goods is equitable, i will probably envy j his looks, or his good health. This problem was alluded to by Kolm (1972). A well-meaning economist who follows his equity theory to its bitter end will conclude that the beautiful should be disfigured, and the well made sick.

Less obvious disadvantages of the idea of equity require references to Pareto efficiency, the foundation of modern welfare economics. An allocation y is *Pareto superior* to an allocation x if all individuals prefer y to x. (This assumes, of course, a constant set of individuals who are making the judgement.) If y is Pareto superior to x, the move from x to y is a *Pareto move*. An allocation x is *Pareto optimal* if there is no y that is Pareto superior to it.

Several authors (e.g. Kolm 1972) have established that in an economy where there is no production, there exist allocations that are both equitable and Pareto optimal. To find one, start at the equal allocation and move the economy to a competitive equilibrium. By the first fundamental theorem of welfare economics, a competitive equilibrium is Pareto optimal. Since the equilibrium is based on the equal allocation, every individual has the same budget. But if i has the same budget as j, he cannot envy the bundle j buys since he could have bought it himself. So this theorem creates a link between equity and the more traditional, more fundamental notion of Pareto optimality.

But it is a weak link. Pazner and Schmeidler (1974) and Varian (1974) consider an economy with production, where i's utility depends not only on his consumption bundle x_i, but also on the number of hours he works q_i. However, production attributes are non- transferable. If person i is ten times as productive as j, there may be no Pareto optimal distribution of consumption goods and of work hours that is also equitable.

Equity, Fig. 1



Think of an economy of which you are a part and Luciano Pavarotti is a part. You would have to train for 10 lifetimes before you could sing an aria like he does, and therefore there may be no possibility of arriving at an allocation of consumption and work effort among all that is both equitable and Pareto optimal.

Various possible solutions to this quandary have been suggested (e.g. in Pazner 1976, and Pazner and Schmeidler 1978). For instance, consider an economy where 'everybody shares an equal property right in everybody's time'. This may lead to the existence of allocations that are both equitable and optimal, but it makes Pavarotti a slave to everyone who is less gifted. Or, as another possible solution, consider an *egalitarian equivalent* allocation. This is one such that the utility distribution it produces could be generated by a theoretical economy in which all consumers are assigned identical consumption bundles. Pazner and Schmeidler (1978) show that egalitarian equivalent allocations that are also Pareto optimal exist, even in economies with production. But this idea is also unworkable; it is simply too airy.

Turn back to an economy without production. It is true that there will exist, under general assumptions, allocations that are both equitable and Pareto optimal in the pure exchange economy. But Feldman and Kirman (1974) show two disturbing facts: First, even if traders start at the equal allocation, and they make a Pareto move to the core (the solution set for frictionless barter), they may end up at an inequitable allocation. Second, if traders start at an equitable allocation, and make a Pareto move to a competitive equilibrium they may end up at an allocation where someone envies someone else. The 'green sickness' springs up where once there was equity.

The Edgeworth box diagram below illustrates the second possibility. In Fig. 1, x_{11} and x_{12} represent quantities of goods 1 and 2 belonging to trader I; x_{J1} and x_{J2} represent quantities belonging to J. Also, i_1 and i_2 are two of trader I's indifference curves: j_1 and j_2 and two of trader J's indifference curves; $w = (w_i, \ w_j)$ is the initial

allocation; $w^{-1} = (w_j, w_i)$ is the allocation which switches the bundles between I and J. Note that w^{-1} is found by reflecting w through the centre of the box. Now w is equitable since the indifference curves through it pass above w^{-1}, and the move from w to x is a competitive equilibrium trade that makes both better off. But $x = (x_i, x_j)$ is not equitable, since i_2 passes below $x^{-1} = (x_j, x_i)$, which means that trader I envies J when they are at x.

In an interesting extension of the Feldman and Kirman result, Goldman and Sussangkarn (1978) show with generality that in 2 person, 2 good exchange economies there exist allocations x such that (a) x is equitable in the non-envy sense but (b) x is not Pareto optimal and (c) every y which is Pareto superior to x is inequitable! This is formal proof of Johnson's assertion (*The Rambler*, No.183) that 'envy is almost the only vice which is practicable at all times, and in every place; the only passion which can never lie quiet from want of irritation'.

The concept of equity as non-envy is still alive among prominent economists; for instance, Baumol (1982) applies non-envy to an analysis of rationing. This in spite of the fact that recent history suggests the average man fares better under regimes that are less committed to elimination of envy through redistribution of goods, and in spite of the serious theoretical objections raised to the concept as outlined above. Should we care about equity? The temptation to pronounce judgement on what is equitable and what is not may be irresistible. But economic theory suggests that the pursuit of equity in the sense of non-envy will lead to some peculiar and unpalatable results.

See Also

► Fairness

Bibliography

Baumol, W. 1982. Applied fairness theory and rationing policy. *American Economic Review* 72(4): 639–651.
Feldman, A., and A. Kirman. 1974. Fairness and envy. *American Economic Review* 64(6): 995–1005.
Foley, D. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7(1): 45–98.
Goldman, S., and C. Sussangkarn. 1978. The concept of fairness. *Journal of Economic Theory* 19(1): 210–216.
Kolm, S.-C. 1972. *Justice et équité*. Paris: Editions du Centre de la Recherche Scientifique.
Pazner, E. 1976. Recent thinking on economic justice. *Journal of Peace Science*.
Pazner, E., and D. Schmeidler. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41(3): 441–443.
Pazner, E., and D. Schmeidler. 1978. Egalitarian equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92(4): 671–687.
Schmeidler, D., and K. Vind. 1972. Fair net trades. *Econometrica* 40(4): 637–642.
Varian, H. 1974. Equity, envy and efficiency. *Journal of Economic Theory* 9(1): 63–91.

Equivalence Scales

Arthur Lewbel and Krishna Pendakur

Abstract

An equivalence scale is a measure of the cost of living of a household of a given size and demographic composition, relative to the cost of living of a reference household (usually a single adult), when both households attain the same level of utility or standard of living. Equivalence scales are difficult to construct because household utility cannot be directly measured, which results in economic identification problems. Applications of equivalence scales include measurement of social welfare, economic inequality, poverty, and costs of children; indexing payments for social benefits, life insurance, alimony, and legal compensation for wrongful death.

Keywords

Consumer expenditure; Engel scales; Equivalence scales; Happiness, economics of; Interpersonal utility comparisons; Marshallian demand functions; Neuroeconomics; Poverty lines; Revealed preference theory; Rothbard scales; Shephard's Lemma; Wellbeing

JEL Classifications
D12

History

Providing two different households with the same standard of living, making them equally well off, requires some definition of well-being. In the early literature on equivalence scales, a household's well-being was defined in terms of needs, such as having a nutritionally adequate diet.

Engel (1895) observed that a household's food expenditures are an increasing function of income and of family size, but that richer households tend to spend a smaller share of their total budget on food than poorer households. He therefore proposed that this food budget share could be a measure of a household's welfare or standard of living. The resulting Engel equivalence scale is defined as the ratio of incomes of two different sized households that have the same food budget share. This is essentially the method used by the US Census Bureau to measure poverty. The bureau first defines the poverty line for a typical household as three times the cost of a nutritionally adequate diet, then uses food shares (Engel scales) to derive comparable poverty lines for households of different sizes and compositions, and finally adjusts the results annually by the consumer price index to account for inflation (see Fisher 1997).

Given two households that differ only in their number or age distribution of children, Rothbarth (1943) equivalence scales are similar to Engel scales. They can be defined as the ratio of incomes of the two households when each household purchases the same quantity of some good that is only consumed by adults, such as alcohol, tobacco, or adult clothing.

Modern equivalence scales measure well-being in terms of utility, using cost (expenditure) functions estimated from consumer demand data via revealed preference theory. Engel or Rothbarth scales are equivalent to valid cost function based equivalence scales only under strong restrictions regarding the dependence of demand functions on characteristics such as age and family size, and on the links between demand functions and utility for these different household types.

One strand of the equivalence scale literature focuses on the former issue, and so deals primarily with the empirical question of how best to model the dependence of household Marshallian demand functions on demographic characteristics. Examples are Sydenstricker and King (1921), Prais and Houthakker (1955), and Barten (1964) scales, in which a different Engel type scale is constructed for every good people purchase, roughly corresponding to a different economies of scale measure for each good. Other examples are Gorman's (1976) general linear technologies, Lewbel's (1985) modifying functions, and Pendakur's (1999) shape invariance.

The second, closely related literature, focuses on the joint restrictions on both preferences and interpersonal comparability of utility required for measuring the relative costs of providing one household with the same utility level as another. Examples include Jorgenson and Slesnick (1987), Lewbel (1989), Blackorby and Donaldson (1993), and Donaldson and Pendakur (2004; 2006).

Definition

Consider a consumer (an individual or a household) with a vector of demographic characteristics $\mathbf{z}$ and nominal total expenditures x that faces the M vector $\mathbf{p}$ of prices of M different goods. The consumer chooses a bundle of goods to maximize utility given a linear budget constraint. Define the cost (expenditure) function $x = C(\mathbf{p}, u, \mathbf{z})$ which equals the minimum expenditure required for a consumer with characteristics $\mathbf{z}$ to attain utility level u when facing prices $\mathbf{p}$. $C(\mathbf{p}, u, \mathbf{z})$ is a conditional cost function in the sense of Pollak (1989) because it gives the expenditure necessary to attain a utility level u, conditional on the consumer having characteristics $\mathbf{z}$.

Equivalence scales relate the expenditures of a consumer with characteristics $\mathbf{z}$ to a consumer with a reference vector of characteristics $\bar{\mathbf{z}}$. The reference vector of characteristics may describe, for example, a single, medically healthy, middle-

aged childless man. The equivalence scale is defined by $D(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, u, \mathbf{z})/C(\mathbf{p}, u, \mathbf{z})$. Equivalent-expenditure $X(\mathbf{p}, x, \bar{\mathbf{z}})$ is defined as the expenditure level needed to bring the well-being of a reference household to the level of well-being of a household with characteristics $\mathbf{z}$, so $X(\mathbf{p}, x, \mathbf{z}) = x/D(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, u, \bar{\mathbf{z}})$ where u is replaced by the indirect utility function, that is, $x = C(\mathbf{p}, u, \mathbf{z})$ solved for u.

Identification

In economics, a parameter is said to be 'identified' if its numerical value can be determined given enough observable data. Here we show why identification of equivalence scales requires either strong untestable assumptions regarding preferences or unusual types of data. Equivalence scales depend on utility, which cannot be directly observed and so must be inferred from consumer demand data, that is, from the quantities that consumers buy of different goods in varying price regimes and at various income levels. The observable (Marshallian) demand functions for goods derived from a conditional cost function $C(\mathbf{p}, u, \mathbf{z})$ are the same as those obtained from $C(\mathbf{p}, \varphi(u, \mathbf{z}), \mathbf{z})$ for any function $\varphi(u, \mathbf{z})$ that is strictly monotonically increasing in u. By revealed preference theory, demand data identifies the shape and ranking of a consumer's indifference curves over bundles of goods, but not the actual utility level associated with each indifference curve. Changing φ(u, $\mathbf{z}$) just changes the utility level associated with each indifference curve.

Therefore, given any $C(\mathbf{p}, u, \mathbf{z})$ derived from demand data, the consumer's true cost of attaining a utility level u is $C(\mathbf{p}, \varphi(u, \mathbf{z}), \mathbf{z})$ for some unknown function 9, so true equivalence scales are $D(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, \varphi(u, \mathbf{z}), \mathbf{z})/C(\mathbf{p}, \varphi(u, \bar{\mathbf{z}}), \bar{\mathbf{z}})$. This is the source of equivalence scale non-identification. We cannot identify $D(\mathbf{p}, u, \mathbf{z})$ because the change from z to $\mathbf{z}$ has an unobservable affect on D through φ. The problem is that revealed preferences over goods identify one set of indifference curves for households of type $\mathbf{z}$ and another set for households of type $\mathbf{z}$, but we have no way of observing which indifference curve of type $\bar{\mathbf{z}}$ yields the same level of utility as any given indifference curve of type $\bar{\mathbf{z}}$.

Given only goods demand data, Blundell and Lewbel (1991) show that changes in equivalence scales that result from price changes can be identified, but the levels of equivalence scales are completely unidentified, because for any cost function C and any positive number d, there exists a $\varphi(u, \mathbf{z})$ function that makes $D(\mathbf{p}, u, \mathbf{z}) = \mathrm{d}$. Changes in D resulting from price changes can be identified because the ratio $D(\mathbf{p_1}, u, \mathbf{z})/D(\mathbf{p_0}, u, \mathbf{z})$ equals a ratio of ordinary identifiable cost of living (inflation) indices.

Identification of equivalence scales therefore requires either additional information or untestable assumptions regarding preferences over characteristics $\mathbf{z}$ and hence regarding φ. There are also other identification issues associated with equivalence scales. For example, different members of a household may have different standards of living, so a single level of utility that applies to the entire household to be compared or equated to anything may simply not exist. Lewbel (1997) lists additional equivalence scale identification issues.

Identification from Demand Data

Let w^j be the fraction of total expenditures a household spends on the jth good (its budget share) and let $\mathbf{w}$ be the vector of budget shares of all purchased goods. Shephard's Lemma states that $\mathbf{w} = \omega(\mathbf{p}, u, \mathbf{z}) = \nabla_{\ln p} \ln C(\mathbf{p}, u, \mathbf{z})$, the price elasticity of cost. Let $w_f = \omega_f(\mathbf{p}, u, \mathbf{z})$ indicate the food equation. Engel's method notes that since ω_f is monotonically declining in utility u, w_f may be taken as an indicator of well-being. If, in addition, w_f indicates the same level of well-being for all household types $\mathbf{z}$, then the expenditure levels which equate the food share w_f, across household types are the equivalent-expenditure function, whose ratios give the equivalence scale. Monotonicity of ω_f in u is observable, but the second restriction concerning utility levels for different types of households refers to φ and so is not testable.

The Rothbarth approach is similar. Let $q_a = h_a(\mathbf{p}, u, \mathbf{z})$ indicate the quantity demanded

for a good consumed only by adults, such as alcohol. If h_a is increasing in utility (a testable restriction), q_a may be taken as an indicator of the well-being of adult household members. If, in addition, q_a indicates the same level of adult wellbeing for adults living in all types of households (untestable), then the expenditure levels which equate q_a across household types are the equivalent-expenditure function, whose ratios again give the (Rothbarth) equivalence scale.

Lewbel (1989) and Blackorby and Donaldson (1993) consider the case where the equivalence scale function is independent of utility, which they call 'independence of base' (IB) and 'equivalence-scale exactness' (ESE), respectively. In this case there is a function Δ such that $D(\mathbf{p}, u, \mathbf{z}) = \Delta(\mathbf{p}, \mathbf{z})$ and $C(\mathbf{p}, u, \mathbf{z}) = C(\mathbf{p}, u, \bar{\mathbf{z}})\Delta(\mathbf{p}, \mathbf{z})$. The special case where $D(\mathbf{p}, u, \mathbf{z})$ is also independent of $\mathbf{p}$ yields Engel scales.

Given IB/ESE, Shephard's Lemma implies that $\omega(\mathbf{p}, u, \mathbf{z}) = \omega(\mathbf{p}, u, \bar{\mathbf{z}}) + \mathbf{n}(\mathbf{p}, \mathbf{z})$, where $\mathbf{n}(\mathbf{p}, \mathbf{z}) = \nabla_{\ln \mathbf{p}} \ln \Delta(\mathbf{p}, \mathbf{z})$. Since households with the same equivalent expenditure have the same utility, and since in this case, equivalent expenditure is given by $x/\mathbf{A}(\mathbf{p}, \mathbf{z})$, we may write the relation as $\mathbf{w}(\mathbf{p},x,\mathbf{z}) = \mathbf{w}(\mathbf{p},x/\Delta(\mathbf{p}, \mathbf{z}), \bar{\mathbf{z}}) + \mathbf{n}(\mathbf{p}, \mathbf{z})$, where $\mathbf{w}(\bullet)$ is the Marshallian budget share vector. Here, $\Delta(\mathbf{p}, \mathbf{z})$ 'shrinks' the budget share functions in the expenditure direction, and the amount of 'shrinkage' identifies the equivalence scale. Pendakur (1999) shows that this 'shape invariance' expression equals the testable implications required for IB/ESE. The untestable restriction, which uniquely defines $\varphi(u, \mathbf{z})$ (up to transformations of u that do not depend on $\mathbf{z}$) is that all households with the same value of $x/\Delta(\mathbf{p}, \mathbf{z})$ have the same level of utility. Blackorby and Donaldson (1993) show when cost functional forms uniquely identify IB/ESE. Donaldson and Pendakur (2004, 2006) consider identification for equivalence scales with more general functional forms.

Other Sources of Identification

Equivalence scale identification depends on how we define utility or well-being. Identification is not a problem if what we mean by making households equally well off refers to some observable characteristic such as nutritional adequacy of diet. As an alternative to revealed preference, identification may be based on surveys that ask respondents to either report their happiness (and hence utility) on some ordinal scale, or ask, based on introspection, how their utility or costs would change in response to changes in household characteristics. An early example is Kapteyn and Van Praag (1976), who estimate equivalence scales based on surveys where households rank income levels as 'excellent', 'sufficient', and so on. Identification requires comparability of these ordinal utility measures across consumers. Happiness studies by psychologists and experimental economists may prove useful for validating these types of subjective responses regarding utility, especially with recent neuroeconomic results measuring brain activity associated with pleasure, regret, and economic decision-making (see, for example, McFadden 2005).

Another possible source of identification is when consumers can choose $\mathbf{z}$, and we can collect information relevant to these choices. Assuming $\mathbf{z}$ is chosen to maximize utility can provide information about how utility varies with $\mathbf{z}$, and hence may restrict the set of possible φ transformations. With enough information regarding how $\mathbf{z}$ is chosen one could identify 'unconditional' cost or utility functions over both goods and $\mathbf{z}$ and thereby identify the dependence of φ on $\mathbf{z}$. Pollak (1989) refers to the use of unconditional versus conditional data to calculate the cost of demographic changes as 'situation comparisons' versus 'welfare comparisons'.

Traditional equivalence scales assign a single level of utility to a household, implicitly assuming that all household members have the same utility level and hence ignoring the effects of the within-household distribution of resources. Features of this intra-household allocation of resources can be identified and estimated with demand data. Given the indifference curves and resource shares of each household member, instead of trying to calculate the cost of making an individual as well off as a household, one may instead calculate the cost of putting the individual on the same indifference

curve when living alone that he attained as a member of a household. Whereas the former calculation requires a welfare comparison, the latter calculation only involves comparing the same individual in two different price and income environments. Browning, Chiappori and Lewbel (2006) call this type of comparison an 'indifference scale', and provide one set of conditions under which such scales can be non-parametrically identified.

Applications of Equivalence Scales

Equivalent expenditures and equivalence scales may be used for social evaluation, for example, inequality and poverty analysis. Given an equivalence scale, *di,* and household expenditure, *xi*, for each person *i* in a population, one constructs equivalent expenditure for each person: $x_i^e = x_i/d_i$. Expenditure data are observed at the level of the household, but x_i^e is constructed for each individual. By construction, the population distribution of equivalent expenditures is equivalent in welfare terms to the actual distribution of expenditures across households. Therefore, one can use this 'as if' distribution for constructing population measures of poverty or inequality, or for calculating the welfare implications of tax and transfer programmes.

Equivalence scales can also be used to calibrate social benefits payments and poverty lines. For example, if the social benefit rate (or poverty line) $\bar{x}$ is agreed upon for a single household type, for example, a single childless adult, then one could use equivalence scales to set rates for other household types $\mathbf{z}$ as $D(\mathbf{p}, u, \mathbf{z})\bar{x}$ where u is the utility level of the reference type with expenditures $\bar{x}$. Some statistical agencies flow information in the other direction: poverty lines are constructed for each household type, which can then be use to construct an implicit 'poverty relative' equivalence scale. If scales are IB/ESE, this provides enough information to identify equivalence scales for all households.

Other applications of equivalence scales are for life insurance, alimony, and wrongful death calculations (see Lewbel 2003), and for indirectly measuring the cost of children based on equivalence scales for households of different sizes.

See Also

- ▶ Consumer Expenditure
- ▶ Cost Functions
- ▶ Demand Theory
- ▶ Engel Curve
- ▶ Engel's Law
- ▶ Hicksian and Marshallian Demands
- ▶ Identification
- ▶ Inequality (Measurement)
- ▶ Interpersonal Utility Comparisons
- ▶ Poverty Lines
- ▶ Welfare Economics

Bibliography

Barten, A.P. 1964. Family composition, prices, and expenditure patterns. In *Econometric analysis for national planning: 16th symposium of the colston society*, ed. P. Hart, L. Mills, and J.K. Whitaker. London: Butterworth.

Blackorby, C., and D. Donaldson. 1993. Adult-equivalence scales and the economic implementation of interpersonal comparisons of well-being. *Social Choice and Welfare* 10: 335–361.

Blundell, R.W., and A. Lewbel. 1991. The information content of equivalence scales. *Journal of Econometrics* 50: 49–68.

Browning, M., P.-A, Chiappori., and A. Lewbel. 2006. *Estimating consumption economies of scale, adult equivalence scales, and household bargaining power*, Working Paper No. 588. Boston: Boston College.

Donaldson, D., and K. Pendakur. 2004. Equivalent-expenditure functions and expenditure-dependent equivalence scales. *Journal of Public Economics* 88: 175–208.

Donaldson, D., and K. Pendakur. 2006. The identification of fixed costs from consumer behaviour. *Journal of Business and Economic Statistics* 24: 255–265.

Engel, E. 1895. Die Lebenskosten Belgischer Arbeiter-Familien Fruher and jetzt. *International Statistical Institute Bulletin* 9: 1–74.

Fisher, G.M. 1997. *The development and history of the U.S. poverty thresholds – A brief overview*. Washington, DC: Newsletter of the Government Statistics Section and the Social Statistics Section of the American Statistical Association.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis: Proceedings of the 1975 AUTE conference*, ed. M. Sheffield, J. Artis, and A.R. Nobay. Cambridge: Cambridge University Press.

Jackson, C.A. 1968. *Revised equivalence scale for estimating equivalent incomes or budget cost by family type*, Bulletin 1570–2. Washington, DC: U.S. Bureau of Labor Statistics.

E

Jorgenson, D.W., and D.T. Slesnick. 1987. Aggregate consumer behavior and household equivalence scales. *Journal of Business and Economic Statistics* 5: 219–232.
Kapteyn, A., and B. Van Praag. 1976. A new approach to the construction of family equivalence scales. *European Economic Review* 7: 313–335.
Lewbel, A. 1985. A unified approach to incorporating demographic or other effects into demand systems. *Review of Economic Studies* 52: 1–18.
Lewbel, A. 1989. Household equivalence scales and welfare comparisons. *Journal of Public Economics* 39: 377–391.
Lewbel, A. 1997. Consumer demand systems and household equivalence scales. In *Handbook of Applied Econometrics,* vol. 2. *Microeconomics*, ed. M.H. Pesaran and P. Schmidt. Oxford: Blackwell.
Lewbel, A. 2003. Calculating compensation in cases of wrongful death. *Journal of Econometrics* 113: 115–128.
McFadden, D. 2005. *The new science of pleasure, consumer behavior and the measurement of well-being*. Frisch Lecture. London: Econometric Society World Congress.
Pendakur, K. 1999. Estimates and tests of base-independent equivalence scales. *Journal of Econometrics* 88: 1–40.
Pollak, R.A. 1989. *The theory of the cost of living index*. New York: Oxford University Press.
Prais, S.J., and H.S. Houthakker. 1955. *The analysis of family budgets*. Cambridge: Cambridge University Press.
Rothbarth, E. 1943. Note on a method of determining equivalent income for families of different composition. In *War-time pattern of saving and spending*, ed. C. Madge. Cambridge: Cambridge University Press.
Sydenstricker, E., and W.I. King. 1921. The measurement of the relative economic status of families. *Quarterly Publication of the American Statistical Association* 17: 842–857.

Ergodic Theory

William Parry

To begin in the middle; for that is where ergodic theory started, in the middle of the development of statistical mechanics, with the solution, by von Neumann and Birkhoff, of the problem of identifying space averages with time averages. This problem can be formulated as follows: If $x_t(-\infty < t < \infty)$ represents the trajectory (orbit) passing through the point $x = x_0$ at time $t = 0$ of a *conservative dynamical system*, when can one make the identification

$$(*)\ \lim_{T\to\infty} (1/T)\int_0^T f(x_t)\mathrm{d}t = \int_\Omega f\,\mathrm{d}m/m(\Omega)$$

for suitable functions defined on the *phase space* Ω of the system?

There are many things to be explained here. For example one might imagine a 'large' number of particles contained in a box, which collide with one another and with the sides of the box according to the usual laws of elastic collision. Each of these particles has three coordinates of position and three coordinates of velocity so that the state of the system is describable by $6n$ coordinates if n is the number of particles. Newtonian laws, of course, provide a history and future for each of these points in $6n$ dimensional space. The same laws imply the law of conservation of energy, so that in principle dynamical systems may be studied with the assumption that energy is constant for each trajectory of a conservative system. Thus in $(^*)$ we take the phase space Ω to be that hypersurface of $6n$ dimensional space where the total energy has a given (constant) value, and m is the hypersurface volume (measure) associated with the Liouville invariant volume whose existence is guaranteed by the conservativity of the system. In general $m(\Omega)$ is a finite quantity.

The left-hand side of $(^*)$ is the time average along a trajectory for a function (observable) f and the right-hand side is the phase or space average.

Von Neumann proved a mean convergence version of $(^*)$ and shortly after G. D. Birkhoff proved $(^*)$ as stated, for *almost all* states, in both cases under the assumption that the system (restricted to Ω) is *ergodic*, a notion, we shall explain presently. (Cf. von Neumann 1932a; Birkhoff 1931.) It was soon realized that both versions of $(^*)$ (the ergodic theorems) could be formulated and proved in a more abstract setting and indeed one can say that this abstraction and the subsequent mathematics thereby generated is ergodic theory proper.

Let (Ω, m) represent an abstract space with a finite measure. (There is no loss in generality in assuming $m(\Omega) = 1$, as we shall do.) Let T_t represent a family of transformations indexed by time (in various contexts, the real numbers, the integers) such that $T_{t+s} = T_t \circ T_s$. Assume that this family is measurepreserving ($mT_tB = mB$, for all 'measurable' sets). The study of T_t as t varies through its index set, provides a model for an evolutionary system, such as the dynamics in phase space described earlier, in which measure (volume) is preserved. The system is said to be *ergodic* if Ω cannot be decomposed into two disjoint invariant measurable sets

$$A, B(A \cup B = \Omega, A \cap B = \varnothing, T_tA = A, T_tB = B \text{ all } t)$$

of positive measure.

In a strict sense, the time-average space-average problem was not *solved* by von Neumann and Birkhoff, as far as the classical dynamical system given at the outset is concerned, for the question of whether this system *is* ergodic was left open and it is only recently (Sinai 1963) that progress has been made in this direction.

Most workers in ergodic theory concern themselves with measure-preserving transformations T_t indexed by the integers, so that with $T_1 = T$, T_t is the iteration of T repeated t times. Results in this context invariably lead to results for real continuous time.

Having freed itself from a particular (albeit important) dynamical system, ergodic theory or more particularly the theory of measure-preserving transformations began to encounter a rich diversity of problems:

(i) When does a measurable transformation, non-singular with respect to a given measure, preserve an equivalent finite (or even σ-finite) measure?
(ii) Are there analogues of the ergodic theorems for Markov processes?
(iii) Where do we find examples of measure-preserving (or non-singular) transformations in other branches? If they are non-singular answer question (i). If they are measure-preserving are they ergodic? If so, interpret the ergodic theorems for them.
(iv) Is it possible to (at least partially) classify the myriad examples coming from other branches of mathematics?

One should notice that in posing these problems ergodic theory became a *global* analysis in two senses: The phase space dynamical system described at the beginning of this entry is global in that *all* solutions of a differential equation are involved. Ergodic theory then moves on to treat all other problems having a dynamical character in which an invariant measure appears.

Concerning (ii) one should note that a measure-preserving transformation T gives rise to an isometric operator $Lf = f^\circ\ T$ on various Banach spaces, the most important being $L^1(m)$. In a similar way a Markov process gives rise to a semi-group of positive contractions. For such operators there is a variety of ergodic theorems generalizing the classical results of Birkhoff and von Neumann. As an example there is the powerful general ergodic theorem (Chacon and Ornstein 1960): If L is a positive contraction on $L^1(m)$ and f, $g \in L^1(m)$ then

$$\sum_{k=0}^{n} L^n f \Big/ \sum_{k=0}^{n} L^n g$$

converges almost everywhere on the set where the denominator is persistently positive.

Here we have an instance of ergodic theory providing a powerful tool for statistics. This should hardly be surprising, however, as even the classical Birkhoff ergodic theorem has an immediate impact on stochastic processes, for one can always associate a measure-preserving transformation with, say, a sequence of independent and identically distributed random variables in such a way that the strong law of large numbers is an easy corollary of Birkhoff's theorem.

Markov and other stochastic processes have played and continue to play a central role in the development of ergodic theory. In recent years a modelling procedure for understanding hyperbolic dynamical systems based on Markov chains

has led to profound results in the area of differentiable statistical mechanics. Thus statistical ideas are exchanged, measure for measure, with those of ergodic theory.

Concerning (iii) here are some examples:

(a) *An 'irrational flow'*. Here $\Omega = \{(z, w)$: z, w complex $| z | = | w | = 1\}$, $T_t(z, w) = (e^{2\pi i\alpha t}z, e^{2\pi i\beta t}w)$, α, β are real with α, β irrational. m is an ordinary Lebesgue measure.

(b) *A skew product*. Here (Ω, m) is the same as in (a).

$$T(z, w) = \left(e^{2\pi i\alpha}z, zw\right), \quad \alpha \, irrational.$$

(c) *An automorphism of a torus*. Again (Ω, m) is the same as in (a).

$$T(z, w) = \left(z^2w, zw\right).$$

(d) *A translation of a homogeneous space*. G is a locally compact Lie group and H is a closed subgroup such that the homogeneous space $G/H = \{gH$: $g \in G\}$ is compact. $\Omega = G/H$ and m is a Haar measure. The transformation T is defined as a translation.

$$T(g, H) = agH$$

for a given element $a \in G$.

(e) *A geodesic flow*. Here we consider an n-dimensional Riemannian manifold M with unit length tangent vectors υ located at points of M. Such a vector υ defines a unique geodesic curve on M. Ω is the totality of such υ and $T_t\upsilon$ is the unit tangent vector obtained by allowing υ to flow along its geodesic at unit speed after time t. The measure m may be taken to be the natural one associated with Liouville's measure.

(f) *A Hamiltonian dynamical system*. Instead of defining this we mention that (e) above and the n particle phase space system at the beginning of this article are both examples of such a system.

(g) *The evolutionary shift associated with a Markov chain or more particularly of a Bernoulli (independent) sequence of trials*.

For the Bernoulli case Ω consists of points $w = \{w_n\}_{-\infty}^{\infty}$ where w_n represents the outcome of an experiment (heads or tails, for example, in the tossing of a coin) at time n. m is the probability which guarantees the independence of these trials, and T is the shift in time $Tw = w'$ where $w'_n = w_{n+1}$.

(h) *A stationary Gaussian (normal process)*.

(i) *The continued fraction transformation*. Here Ω consists of the irrational numbers between 0 and 1. m is 'Gauss's' measure whose density is $1/\log 2\,(1 + x)$ and $Tx = 1/x \bmod 1$.

An alternative account of ergodic theory, which admittedly ignores the history of the subject, could be given which is based on the above examples (and many others). It would motivate the subject by the questions: What do these examples have in common? What concepts underlie them? However, only a posteriori would these questions lose their artificiality.

The first four examples (a), (b), (c) and (d), all arise from algebraic or homogeneous space structures and even (e) falls into this category under certain conditions on the curvature of the manifold. In general, (e) arises from differential geometry. The Bernoulli example (g) (or more generally a Markov chain) arises from probability theory as does (h). The example (i) occurs in the study of continued fractions.

These examples (under suitable conditions) are flows and transformations which display varying degrees of ergodicity or *mixing* and ergodic theoretical techniques reveal important information about them. For example, in Furstenberg, (1961) (b) was used to give a proof of the famous theorem of Weyl that $\alpha n^2 + \beta n + \gamma \bmod 1$ is uniformly distributed in the unit interval [0, 1] as n varies (as long as α or β is irrational). The example (c) was closely analysed as a prototype of *hyperbolicity* prior to the development of Anosov and Axiom A dynamical systems. The examples covered by (e) (and the related *horocycle flows*) are central to the study of hyperbolic geometry and to the theory of unitary representations of semi-simple Lie groups. The examples (g), (h) provide the most important classes of stationary stochastic processes and are intimately related to Brownian

motion. Example (i) is of vital importance in number theory.

Question (iv) was first approached (von Neumann 1932b; Halmos and von Neumann 1942) using spectral techniques. Two measure-preserving transformations *S, T* are said to be (spatially) *isomorphic* if there is an invertible measure-preserving transformation ϕ between their respective spaces such that $\phi S = T\phi$ a.e. (almost everywhere). Isomorphism implies that the unitary spectral characteristics are indistinguishable (i.e. *spectral* equivalence), but not vice versa. The main result obtained characterized all ergodic measure-preserving transformations with a pure point spectrum. For such transformations *S, T* identity of point spectrum implies spatial isomorphism and such transformations are (isomorphically) precisely the ergodic translations of compact metric abelian groups.

A similar theory was developed in Abramov (1962) for so-called transformations with quasi-discrete spectrum. Example (b) provides an example of this type of transformation. They had been studied earlier by Anzai. The works of Auslander, Green and Hahn (1963) and Parry (1971) provide further developments in this direction. A completely analogous theory is modelled on the 'rigid' examples of *nilflows* and *unipotent affines* on nil manifolds. The rigidity here refers to the phenomenon of measure isomorphisms *necessarily* being algebraic in character. The most recent work concerning rigidity in ergodic theory (Ratner 1982) finds this, and related phenomena, in horocycle flows.

So far we have given a condensed account of only one strand in isomorphism theory. The most active work has occurred in connection with examples of an entirely different and *random* character.

This work began with the problem of deciding whether two Bernoulli shifts (which are necessarily spectrally isomorphic) are spatially isomorphic. The first breakthrough occurred with Kolmogorov's introduction of entropy theory into the subject (Kolmogorov 1958). As modified in (Sinai 1959) entropy is a numerical invariant of isomorphism (i.e. if *S, T* are isomorphic then their entropies *h(S), h(T)* coincide). This fact provides a multitude of Bernoulli transformations which are not isomorphic. The basic ideas originate with Shannon and McMillan, but they required significant adaptation before they could be used in ergodic theory. The new entropy theory developed apace in the hands of, principally, Russian mathematicians in the 1960s and received its biggest impetus from the American mathematician Ornstein, who in 1968 proved that two Bernoulli transformations with the same entropy are isomorphic (cf. Ornstein 1970). From that time the subject has grown exponentially, with ever more transformations shown to be (isomorphic to) Bernoulli transformations. Such transformations have to have (to say the least) positive entropy and their *intrinsic* random character is in marked contrast to the rigid examples referred to earlier which are *deterministic* (with zero entropy).

Entropy plays very little role in the looser classification theory which allows velocities (along trajectories) to vary. There are continuous (real) time and discrete versions of this theory and as early as 1943 Kakutani had conjectured that all ergodic systems are *Kakutani equivalent*, using the current nomenclature for this loose equivalence (Kakutani 1943). Although this conjecture turned out to be false (in fact entropy ensures the existence of at least three Kakutani inequivalent systems), Feldman (1976) and Katok (1977) showed that remarkably dissimilar systems are equivalent according to this notion. In Ornstein and Weiss (1984) it is shown that a modification of Kakutani's conjecture is true. In this connection a grand theory of equivalence relations in ergodic theory has been developed in Rudolph (1984). This is ergodic theory with its head in the clouds.

From a more earthly point of view ergodic theory in the 1960s, through the developments of entropy theory and stimulated by Anosov (1967) and Smale (1967), began to connect with the newly flourishing field of differentiable dynamical systems.

Examples (c) and (f) are, respectively, prototypes of Anosov diffeomorphisms and flows. Their principal feature here is their global hyperbolic structure. Dynamicists are particularly interested in such systems as they are *structurally stable*, a concept which became something of a

dogma in the 1960s and 1970s, as some mathematicians went so far as to assert that any real and persistent system must be structurally stable. (A structurally stable system is, roughly speaking, one which retains its principal features after a small perturbation.) This important concept was modified by Smale when he introduced Axiom A systems and proved that the latter are Ω-stable (structural stability relative to non-wandering sets). Smale thereby axiomatized a vast category of new dynamical systems and presented us with an approach which unified Anosov systems, gradient like dynamical systems and his so-called 'horse shoes'. For these systems Smale proved his spectral decomposition theorem, which describes the non-wandering set of an Axiom A system in much the same way as one describes the irreducible block behaviour of a non-negative matrix, in the theory of Markov chains (Smale 1970). The *basic* sets of an Axiom A system received further scrutiny in terms of Markov partitions by Sinai (Anosov case) and Bowen (Axiom A case).

Bowen was a key figure in the fruitful convergence of ergodic theory and differentiable dynamical systems because of his profound expertise in both subjects. In a series of papers he provided deep analyses of Axiom A diffeomorphisms and flows (roughly speaking, hyperbolic dynamics) from the point of view of symbolic dynamics and periodic orbits (Bowen 1977). His work connected happily with the direction Ruelle and Sinai were taking in statistical mechanics (Sinai 1972; Ruelle 1978). Together they laid the foundations for statistical mechanics on manifolds.

The subject then has gone full circle to its origins, but on the way it encountered a dazzling variety of iteration problems from other areas, viz. maps of the unit interval (Collet and Eckmann 1980), boundary measures associated with Fuchsian groups (Patterson 1976; Sullivan (1979), analytic maps of the Riemann sphere or complex plane (Rees 1982), to name but three.

As to recent developments in statistical mechanics the one-dimensional lattice gas has received the most attention. Here one considers a shift transformation (as in the case of a Markov chain) initially in the absence of any probability but supplemented with a natural topology which reflects the connectivity of the transformation. Such a shift is called a *topological Markov chain* (or *shift of finite type)*.

Then one considers an action potential describable in terms of a continuous function. Under a stronger (Lipschitz) condition it turns out that there is always a unique shift invariant probability (called an *equilibrium state)* given by a variational principle involving the *pressure* of the potential.

A key tool in this theory is the *transfer* matrix or operator associated with the potential and under suitable (aperiodic and irreducible) conditions, iterations of this operator will force arbitrary probabilities to converge to the equilibrium state.

The one-dimensional lattice gases provide models for simple gases and also for statistical mechanics on manifolds. The results above have analogues (when appropriate conditions are imposed) for differential or even topological dynamical systems (Pesin 1977). Moreover, at least for hyperbolic systems, one can view topological Markov chains with their potentials, equilibrium states, closed orbits, transfer operators and pressures as (in a technical sense) building schemes for these systems.

A recent new area of ergodic theory which stands outside the developments just sketched is concerned with the application of ergodic theory and topological dynamics to combinatorial number theory. The motivation for this recent work was Szemeredi's proof of a conjecture of Erdos and Turan. The conjecture, which emanated from a result of van der Waerden's, states that if $a_1 < a_2 < \cdots$ is an increasing sequence of positive integers and if θ_N denotes the number of these integers less than N, then for every $k > 0$ there is an arithmetic progression in the sequence of length k, as long as $\theta_N/N > \epsilon$ infinitely often (for some $\epsilon > 0$).

Furstenberg (1977) provided an ergodic theoretical proof of this result and of many other results with the same 'flavour'. His technique involved building a sequence $\{x_n\}$ of zeros and

ones ($x_n = 1$ when and only when n is in the sequence), and embedding this sequence in a shift space. The details, which are quite intricate, involve the proof of a multiple recurrence theorem:

If T is a measure-preserving transformation on (X, m) and if $m(A) > 0$ then for every positive integer k

$$m\left(A \cap T^n A \cap T^{2n} A \cap \cdots \cap T^{kn} A\right) > 0$$

for infinitely many integers n.

Although this area is somewhat askew to the other developments outlined above, it needs to be mentioned because of the great research potential it possesses.

Where are the likely growing points for the subject? Here is a list of guesses. Some of them are wild; others are safe; and they are not all of equal weight:

(1) Applications to combinatorial number theory (Furstenberg 1977);
(2) Problems involving a mixture of prime number theory and ergodic theory inspired perhaps by Vinogradov's theorem that $p\alpha$ mod 1 is uniformly distributed when p runs through the primes and α is irrational;
(3) Greater understanding of the connections between the prime number theorem and the prime orbit theorem (Hejhal 1976; Parry and Pollicott 1983);
(4) Further developments of cohomology theory in ergodic theory (Schmidt 1977);
(5) Developments in *restricted* classification theories of processes; in particular a solution of Williams's problem (Williams 1973); in particular a solution of the stochastic version of the theory of Adler and Marcus (1979);
(6) Developments from Ornstein's and Weiss's modified Kakutani problem in the theory of von Neumann algebras;
(7) Which ergodic translations, affines and flows are rigid?
(8) A greater understanding of turbulence (Ruelle and Takens 1971).

See Also

▶ Continuous and Discrete Time Models
▶ Continuous-Time Stochastic Models
▶ Continuous-Time Stochastic Processes

Bibliography

Abramov, L.M. 1962. Metric automorphisms with quasi-discrete spectrum. *Izvestiya Akademii Nauk Seriya Matematicheskaya* 26: 513–530; *American Mathematical Society Translations* 2(39): 37–56.

Adler, R.L., and B. Marcus. 1979. Topological entropy and equivalence of dynamical systems. *Memoirs of the American Mathematical Society* 219: 1–84.

Anosov, D.V. 1967. Geodesic flows on closed Riemannian manifolds with negative curvature. Trudy Matematicheskogo Instituta imeni VA Steklova 90: 1–209; *Proceedings of the Steklov Institute of Mathematics (American Mathematical Society Translations)*, 1969, 1–235.

Auslander, L., L. Green, and F. Hahn. 1963. *Flows on homogeneous spaces*, Annals of mathematics studies, vol. 53. Princeton: Princeton University Press.

Birkhoff, G.D. 1931. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America* 17: 656–660.

Bowen, R. 1977. On axiom A diffeomorphisms. *American Mathematical Society Regional Conference Series* 35: 1–45.

Chacon, R.V., and D.S. Ornstein. 1960. A general ergodic theorem. *Illinois Journal of Mathematics* 4: 153–160.

Collet, P., and J.P. Eckmann. 1980. *Iterated maps on the interval as dynamical systems. Progress in physics*, vol. 1. Boston: Birkhauser.

Feldman, J. 1976. Non-Bernoulli K-automorphisms and a problem of Kakutani. *Israel Journal of Mathematics* 24: 16–37.

Furstenberg, H. 1961. Strict ergodicity and transformations of the torus. *American Journal of Mathematics* 83: 573–601.

Furstenberg, H. 1977. Ergodic behaviour of diagonal measures and a theorem of Szemeredi on arithmetic progressions. *Journal d'analyse mathématique* 31: 2204–2256.

Halmos, P.R., and J. von Neumann. 1942. Operator methods in classical mechanics II. *Annals of Mathematics* 43: 332–350.

Hejhal, D.A. 1976. The Selberg trace formula and the Riemann zeta function. *Duke Mathematical Journal* 43: 441–482.

Kakutani, S. 1943. Induced measure-preserving transformations. *Proceedings of the Imperial Academy of Tokyo* 19: 635–641.

Katok, A. 1977. Monotone equivalence in ergodic theory. *Izvestiya Akademii Nauk Seriya Matematicheskaya* 41: 104–157.

Kolmogorov, A.N. 1958. A new metric invariant of transient dynamical systems and automorphisms of Lebesgue spaces. *Doklady Akademii Nauk SSSR* 119: 8561–8864 (Russian).
Ornstein, D.S. 1970. Bernoulli shifts with the same entropy are isomorphic. *Advances in Mathematics* 4: 337–352.
Ornstein, D.S., and B. Weiss. 1984. Any flow is the orbit factor of any other. *Ergodic Theory and Dynamical Systems* 4: 105–116.
Parry, W. 1971. Metric classifications of ergodic nil flows and unipotent affines. *American Journal of Mathematics* 93: 819–828.
Parry, W., and M. Pollicott. 1983. An analogue of the prime number theorem for closed orbits of axiom A flows. *Annals of Mathematics* 118: 573–591.
Patterson, S.J. 1976. The limit set of a Fuchsian group. *Acta Math* 136: 241–273.
Pesin, J. 1977. Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys* 32(4): 55–114.
Ratner, M. 1982. Rigidity of horocycle flows. *Annals of Mathematics* 115: 597–614.
Rees, M. 1982. *Positive measure sets of ergodic rational maps*. University of Minnesota Mathematics report.
Rudolph, D. 1984. *Restricted orbit equivalence*. Reprinted, Baltimore University of Maryland.
Ruelle, D. 1978. *Thermodynamic formalism*. Reading: Addison-Wesley.
Ruelle, D., and F. Takens. 1971. On the nature of turbulence. *Communications in Mathematical Physics* 20: 167–192.
Schmidt, K. 1977. *Cocycles on ergodic transformation groups*. London: Macmillan.
Sinai, J.G. 1959. On the concept of entropy of a dynamical system. *Doklady Akademii Nauk SSSR* 124: 768–771.
Sinai, J.G. 1963. On the foundations of the ergodic hypothesis for a dynamical system of statistical mechanics. *Doklady Akademii Nauk SSSR* 153: 1261–1264; *Soviet Mathematics - Doklady* 4: 1818–1822, 1963.
Sinai, J.G. 1972. Gibbsian measures in ergodic theory. *Usphehi Matematiceskich Nauk* 27(4): 21–64; *Russian Mathematical Surveys* 27(4): 21–69.
Smale, S. 1967. Differentiable dynamical systems. *Bulletin of the American Mathematical Society* 73: 747–817.
Sullivan, D. 1979. The density at infinity of a discrete group of hyperbolic motions. *Publications mathématiques* 50: 419–450.
Von Neumann, J. 1932a. Proof of the quasi-ergodic hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 18: 70–82.
Von Neumann, J. 1932b. Zur operatoren Methode in der klassischen Mechanik. *Annals of Mathematics* 33: 587–642.
Walters, P. 1973. A variational principle for the pressure of continuous transformations. *American Journal of Mathematics* 97: 937–971.
Williams, R.F. 1973. Classification of subshifts of finite type. *Annals of Mathematics* 88: 120–193.

Ergodicity and Nonergodicity in Economics

Ulrich Horst

Abstract

A random economic system is called ergodic if it tends in probability to a limiting form that is independent of the initial conditions. Breakdown of ergodicity gives rise to path dependence. We illustrate the importance of ergodicity and breakdown thereof in economics by reviewing some work of non-market interactions. This includes microeconomic models of endogenous preference formation, macroeconomics models of economic growth, and models of social interaction.

Keywords

Ergodicity and non-ergodicity in economics; Path dependence; Endogenous preference formation; Ising economy; Gibbs distribution theory; Markov processes; Social interaction

JEL Classifications

D85

A stochastic system is called ergodic if it tends in probability to a limiting form that is independent of the initial conditions. Breakdown of ergodicity gives rise to path dependence. Path-dependent features of economics range from small-scale technical standards to large-scale institutions. Prominent examples include technical standards, such as the 'QWERTY' standard typewriter keyboard and the 'standard gauge' of railway track. Ergodicity and breakdown thereof is of particular relevance to models of social interaction. We illustrate this importance, summarizing some work on endogenous preference formation, dynamic population games, and models of nonmarket interaction.

Endogenous Preference Formation

In his pioneering paper on endogenous preference formation, Föllmer (1974) developed an equilibrium analysis of large exchange economies where the conditional excess demand $z(x^a, p)$ of the agent $a \in \mathbf{A}$ given a price system p is subject to a random shock x^a and where the probabilities $(\pi_a)_{a \in \mathbf{A}}$ governing this randomness have an interactive structure.

In a benchmark model where the *states* x^a are independent across agents, the distribution μ of the vector of states $x = (x^a)_{a \in \mathbf{A}}$ takes the product form $\mu = \Pi_{a \in \mathbf{A}} \pi_a$, and the law of large numbers yields

$$\lim_{n\to\infty} \frac{1}{|\mathrm{A}_n|} \sum_{a \in \mathbf{A}_n} z(x^a, p) = \int z(x^a, p)\mu(dx)\mu - \text{almost surely} \qquad (1)$$

for an increasing sequence of finite populations $\{\mathbf{A}_n\}_{n \in \mathbf{N}}$. Under standard conditions on $z(x^a, \cdot)$ there exists a unique price system p^* for which per capita excess demand is small in economies with many agents, that is, for which

$$\int z(x^a, p^*)\mu(dx^a) = 0. \qquad (2)$$

The assumption of independence of states can be dropped as long as μ is *ergodic*, that is, as long as (1) holds. However, when preferences are interactive, the probabilities π_a specifying the dependence of the individual states on the states of others do not necessarily determine the *joint* distribution μ of all the states. This effect can best be illustrated by means of an 'Ising economy' where the agents are indexed by the two dimensional integer lattice ($\mathbf{A} = \mathbf{Z}^2$), the set of possible states is $\{-1, 1\}$, and where the conditional distribution of agent a's state depends on all the other states x^{-a} only through the states x^b of his four nearest neighbours $b \in N(a) := \{\hat{a} \in \mathbf{A} : |a - \hat{a}| = 1\}$. The distribution also depends on some constant $h \in \mathbf{R}$ which assigns an intrinsic value to private states and on a non-negative quantity J that measures the strength of social interactions. Specifically,

$$\pi_a(x^a; x^{-a}) = \frac{\exp\left\{x^a h + x^a \sum_{b \in N(a)} J x^b\right\}}{\exp\left\{x^a h + x^a \sum_{b \in N(a)} J x^b\right\} + \exp\left\{-x^a h - x^a \sum_{b \in N(a)=1} J x^b\right\}}. \qquad (3)$$

A probability measure μ on $S = \{x = (x^a)_{a \in \mathbf{A}}: x^a \in \{-1, +1\}\}$ is called a *global phase* if its one-dimensional marginal distributions are consistent with the microscopic data given by the individual characteristics $(\pi_a)_{a \in \mathbf{A}}$, that is, if

$$\mu(x^a = \pm 1 | x^{-a}) = \pi_a(\pm 1; |x^{-a}). \qquad (4)$$

An ergodic phase μ can be *equilibrated* if there exist prices p^* for which (2) holds. For independent preferences ($J = 0$) global phases and, hence, equilibrium prices are always determined uniquely by the agents' characteristics. However, if the distribution of states depends only on the states of others ($h = 0$) and the interaction is sufficiently strong, that is, if J exceeds some critical value, two ergodic global phases μ_+ and μ_- exist. In this case aggregate behaviour cannot be inferred from looking at microscopic characteristics alone. Moreover, there is typically no price system that equilibrates both phases *simultaneously*. Thus, randomness in preferences becomes a source of uncertainty about market clearing prices.

Stochastic Strategy Revision in Population Games

The pioneering work by Blume (1993) puts Föllmer's model into a dynamic framework of interactive choice and exploits the link of discrete choice models with Gibbs distribution theory. It is

mainly concerned with the aggregate behaviour in population games of bounded rational play, looking for 'Nash-like play in the aggregate rather than at the level of an individual player'. In Blume's model choice opportunities arise randomly according to individual players' Poisson 'alarm clocks'. When a choice opportunity arises for player $a \in \mathbf{A}$ at time t, his choice x_t^a results in an instantaneous payoff $G(x_t^a, x_t^b)$ from each neighbor $b \in N(a)$ and in a total payoff

$$\sum_{b \in N(a)} G(x_t^a, x_t^b). \tag{5}$$

The conditional probability $\pi_a(x_t^a; x_t^{-a})$ with which player $a \in \mathbf{A}$ selects an action x_t^a at time t, given the current states x_t^{-a} of all the other agents takes the form (3) with $h = \beta\hat{h}$ and $J = \beta\hat{J}$. Here $\beta \geq 0$ specifies the strength of interaction. For $\beta = 0$ the agents choose the actions with equal probability while a best response dynamics corresponds to the limiting case when β tends to infinity. The constants $\hat{h}$ and $\hat{J}$ are determined *endogenously* by the payoff matrix G through (5). Specifically,

$$\pi_a(x_t^a; x_t^{-a}) = \frac{\exp\left\{\beta\left[x_t^a\hat{h} + x_t^a\sum_{b \in N(a)}\hat{J}x_t^b\right]\right\}}{\exp\left\{\beta\left[x_t^a\hat{h} + x_t^a\sum_{b \in N(a)}\hat{J}x_t^b\right]\right\} + \exp\left\{-\beta\left[x_t^a\hat{h} + x_t^a\sum_{b \in N(a)}\hat{J}x_t^b\right]\right\}}.$$

These *flip rates* generate a continuous time Markov process X on S which describes the evolution of the agents' choices through time. A probability measure μ is called an *ergodic measure* for X if the distribution of choices does not change over time and empirical averages converge to a deterministic limit if the initial state is chosen according to μ. The process X is called *ergodic* if it has a unique ergodic measure. It is well known from the theory of interacting particle systems that the set of all ergodic probability measures for X is given by the ergodic global phases corresponding to the local specification (3). As a result, Blume's stochastic strategy revision process is ergodic if $\hat{h} \neq 0$ and $\hat{J} \geq 0$. This is the case if G describes a two person coordination game. Ergodicity breaks down for games with symmetric payoff matrices ($\hat{h} = 0$) when the interaction gets too strong. In this case, the long-run average choice depends on the starting point. The long-run macroscopic behaviour is as unpredictable as equilibrium prices in Föllmer's model by looking at microscopic characteristics only.

Non-ergodic Economic Growth

The evolution of individual choices in Blume (1993) is described by a *continuous* time Markov process with *asynchronous* updating. In local interaction models with *synchronous* updating, the dynamics of individual behaviour is typically described by a Markov chain whose transition operator takes the product form

$$\prod(x_t;\cdot) = \prod_{a \in \mathbf{A}} \pi_a(\cdot; \{x_t^b\}_{b \in (a)}). \tag{6}$$

Thus, the distribution of the state x_{t+1}^a in period $t + 1$ depends on the neighbours' states $\{x_t^b\}_{b \in N(a)}$. in period t. The long-run dynamics of such Markov chains plays an important role in macroeconomic models of economic growth.

The substantial differences in output levels and growth rates across countries have long been a major focus of macroeconomic research. A hallmark of the stochastic growth model pioneered by Brock and Mirman is the convergence of economies with identical preferences and production functions to a common level of aggregate output. Yet many analyses of long-run output movements have concluded that per capita production is not equalizing across countries. To explain this divergence, Durlauf (1993) studies a dynamic model of capital accumulation of an economy with an infinite set $\mathbf{A}$ of interacting companies where local technological externalities

affect the process of production. Each company $a \in \mathbf{A}$ chooses a capital stock sequence $\{K_t^a\}_{t\in}$ that maximizes the present value of future profits, and the technique-specific production functions generate output

$$Y_t^a = f\left(K_{t-1}^a, x_t^a, F\left(x_t^a\right)\right) \tag{7}$$

where $x_t^a \in \{0,1\}$. Technique $x_t^a = 1$ is more productive, but comes at a higher fixed cost: $F(1) > F(0)$. Local technological complementarities affect the production as the distribution of x_t^a depends on the techniques implemented by the nearest neighbours $b \in N(a)$ in the previous period. The dynamics of production technologies is then described by an interactive Markov chain of the form (6). Assuming that past choices of technique 1 improve the current relative productivity of the technique and that the high-productivity state $x_t^a = 1$ for all $a \in \mathbf{A}$ is an equilibrium, Durlauf (1993) shows that the high-productivity state is the only longrun outcome if the complementarities are weak enough: there exists $0 < \theta < 1$ such that

$$\lim_{t\to\infty} \mathrm{P}\left[x_t^a = 1 \middle| x_0^a = 0\right] = 1 \text{ if } \pi_a\left(1; \left\{x_{t-1}^b\right\}_{b\in N(a)}\right) \times \geq \underline{\theta}.$$

Even when one starts with all low-production industries, an economy eventually coordinates on the high-production technology when negative feedbacks from lowproduction technologies are sufficiently weak. Powerful negative complementarities, on the other hand, can generate a non-ergodic growth path. In fact, there exists $0 < \overline{\theta} < \underline{\theta} < 1$, such that

$$\lim_{t\to\infty} \mathrm{P}\left[x_t^a = 1 \middle| x_0^a = 0\right] < 1 \text{ if } \pi_a\left(1; \left\{x_{t-1}^b\right\}_{b\in N(a)}\right) \times \leq \overline{\theta}.$$

If the complementarities are too strong, industries fail to coordinate on highproductivity equilibria, and economies may get trapped in low-productivity equilibria.

Models of Social Interaction – Mean-Field Interaction

Much of the literature on social interactions assumes very special interaction structures such as nearest neighbour interactions as in Blume (1993), or Durlauf (1993) or mean-field interaction. If agents care about the average behaviour throughout the whole population, the analysis is most naturally done in the context of an infinity of agents, as in Brock and Durlauf (2001). These authors analyse aggregate behavioural outcomes when individual utility exhibits social interaction effects. In the simplest setting agents take actions xa from the binary action set $\{-1, +1\}$ and their utilities consists of three components:

$$U^a(x^a, m^a, \varepsilon(x^a)) = u(x^a) + Jx^a m^a + \varepsilon(x^a), \tag{8}$$

Here m^a denotes agent a's expectation about the average choice of all the other agents. The second term in the utility function may thus be viewed as a social utility expressing an agent's desire for conformity ($J > 0$). The quantity $u(x^a)$, on the other hand, represents the private utility associated with a choice while $\varepsilon(x^a)$ is a random utility term independent of other agents' utilities and extreme-value distributed with parameter $\beta > 0$. The extreme-value distribution assumption for the random utility term yields conditional choice probabilities π_a of the form (3) if we replace the dependence of *actual* actions by a dependence on *expected* actions. When agents have homogeneous expectations about the behaviour of others ($m^a \equiv m$), then

$$\pi_a(x^a; m) = \frac{\exp\{\beta(u(x^a) + Jx^a m)\}}{\exp\{\beta(u(1) + Jm)\} + \exp\{\beta(u(-1) - Jm)\}}. \tag{9}$$

In the limit of an infinite economy all uncertainty about the average action vanishes because the agents' choices are conditionally independent given their expectations about aggregate behaviour. The average action is $tanh(\beta h + \beta Jm)$ where

$h = \frac{1}{2}(u(1) - u(-1))$. If the agents have rational expectations the average satisfies the fixed point condition

$$m = \tanh(\beta h + \beta J m). \tag{10}$$

This equation has a unique solution if $h \neq 0$ and β is large enough. For large enough β the uniqueness property breaks down if $h = 0$, in which case (10) has three roots.

Models of Social Interaction – Local and Global Interaction

When agents care about both the average action and the choices of neighbours, the equilibrium analysis becomes more involved. Horst and Scheinkman (2006) provide a general framework for analysing systems of social interactions with an infinite set of locally and globally interacting agents located on an integer lattice (for example, $\mathbf{A} = \mathbf{Z}^2$), continuous action spaces and random preferences. Specifically, they consider utility functions of the form

$$u^a(x, \theta^a) = U\left(x^a, \{x^b\}_{b \in N(a)}, \rho(x), \theta^a\right)$$

where $\rho(x)$ denotes the average choice associated with the action profile x, and the random variables θ^a specify the distribution of taste shocks. While the distinction between local and global interactions is unnecessary for models with finitely many agents, it is important for the analysis of infinite economies. The continuity of the utility functions $u^a(\cdot, \theta^a)$ in the product topology on the configuration space requires, implicitly, that the dependence of an agent's utility function on another agent's action decays sufficiently fast as the distance from that other agent grows. Thus, if preferences depend on average actions, utility functions are typically discontinuous. To overcome this problem, Horst and Scheinkman (2006) separated the local and global impact of an action profile $x = (x^a)_{a \in \mathbf{A}}$ on individual preferences by viewing the average action as an additional parameter, ρ, of a *continuous* utility function on an *extended* state space. The parameter ρ can be seen as the agents' common expectation about the average behaviour. Under standard curvature conditions on U an equilibrium x^ρ exists for any such expectation ρ. If some form of spatial homogeneity prevails and under a weak interaction condition that restricts the influence of an agent's choice on the optimal decisions of others, x^ρ is unique. Furthermore, there exists a unique ρ that coincides with the average action $\rho(x^\rho)$ associated with x^ρ. In this case the agents correctly anticipate the average behaviour, and x^ρ turns out to be the unique equilibrium. The weak interaction condition also guarantees spatial ergodicity: the equilibrium of the infinite system is the limit of equilibria of finite systems when the number of agents grows to infinity; see Horst and Scheinkman (2005) for details.

Dynamic Models of Social Interaction

When dynamic models of social interaction are studied the analysis is often confined to the case of backward-looking myopic dynamics, either as a simple explicit dynamic process with random sequential choice or as an equilibrium selection procedure. Rational expectations equilibria of economies with local interactions are studied in Bisin et al. (2006). While agents interact locally in these models, they are forward-looking. Their choices are optimally based on the past actions in their neighbourhood as well as on their anticipations of the future actions of their neighbours. The resulting population dynamics can be described by an interactive Markov chain of the form (6) but the transition probabilities πa are *endogenously* specified in terms of the agents' policy functions. Bisin et al. (2006) also allow for local and global interactions and combine spatial and temporal ergodicity results. The dynamics on the level of aggregate behaviour is deterministic (spatial ergodicity) and the distribution of individual choices settles down in the long run (temporal ergodicity) when the interaction is weak enough. The analysis, however, is confined to one-sided interactions. It is an open problem to fully embed the theory of social interactions into a dynamics analysis of equilibrium.

See Also

- ▶ Agent-based Models
- ▶ Social Interactions (Empirics)
- ▶ Social Interactions (Theory)
- ▶ Social Multipliers

Bibliography

Bisin, A., U. Horst, and O. Özgür. 2006. Rational expectations equilibria of economies with local interactions. *Journal of Economic Theory* 127: 74–116.

Blume, L. 1993. The statistical mechanics of strategic interactions. *Games and Economic Behavior* 5: 387–424.

Brock, W., and S. Durlauf. 2001. Discrete choice with social interactions. *Review of Economic Studies* 68: 235–260.

Brock, W., and L. Mirman. 1972. Optimal growth under uncertainty: The discounted case. *Journal of Economic Theory* 4: 479–513.

Durlauf, S. 1993. Nonergodic economic growth. *Review of Economic Studies* 60: 349–366.

Föllmer, H. 1974. Random economies with many interacting agents. *Journal of Mathematical Economics* 1: 51–62.

Horst, U., and J. Scheinkman. 2005. A limit theorem for systems of social interactions. Working paper.

Horst, U., and J. Scheinkman. 2006. Equilibria in systems of social interactions. *Journal of Economic Theory.* 127: 74–116.

Erhard, Ludwig (1897–1977)

Ralf Dahrendorf

Erhard was a man who had his moment in history and grasped it. As head of the Economic Department of the administration which preceded the creation of the Federal Republic of Germany, he was the author of the decision to combine the currency reform of 1948 with the abolition of rationing, and of restrictive regulations concerning production, distribution and capital movements. Many have argued that Germany's 'economic miracle' (and not less the political miracle) owes much to these decisions which at the time were regarded as either unrealistic or indefensible by many, including the Occupation Powers.

In a sense, Erhard's life before 1948 was a preparation for this moment, and his career afterwards a continuation of its theme. Born in Fürth in Franconia into a small business family, Erhard studied economics after World War I and joined an economic research institute. His teachers were, on the one hand, Wilhelm Rieger, first director of the Nuremberg Commercial College, and, on the other, Franz Oppenheimer, economist and sociologist in Frankfurt, whose influence on Erhard went much deeper. In the Sixties Erhard described Oppenheimer's importance for him in this way: his own economic policy was in a sense the redirection of Oppenheimer's 'liberal socialism' to 'social liberalism'. During World War II he wrote a memorandum sketching his project for a market economy in ways which left no doubt that he foresaw and wished for the defeat of the Nazis. This was one reason why he was appointed Bavarian Minister of Economic Affairs in 1945, and in 1947, head of the small special unit which prepared the currency reform of 1948. When Konrad Adenauer formed the first Federal Government, Erhard became Minister of Economic Affairs, a post which he held until he succeeded Adenauer as Federal Chancellor in 1963. It was as Economics Minister that Erhard preached and implemented the concept of 'social market economy', a market economy tempered by basic social policies, for which the Federal Republic has become famous. Erhard's Chancellorship was undistinguished; in 1966, his own party, the Christian Democratic Union (CDU) forced him to resign. However, his effect on Germany's economic institutions and the prevailing mould of economic thought is profound and lasting.

References

Caro, M.K. 1965. *Der Volkskanzler – Ludwig Erhard*. Cologne: Kiepenheurer & Witsch.

Lukomski, J.M. 1965. *Ludwig Erhard – Der Mensch und der Politiker*. Dusseldorf: Econ.

Erlich, Alexander (1913–1985)

Diane Flaherty

Keywords

Bukharin, N.I.; Erlich, A.; Industrialization; Preobrazhensky, E.A.; Socialism; Soviet Union, economics in

JEL Classifications

B31

Alexander Erlich was born in St Petersburg on 6 December 1913 and died on 7 January 1985. He moved to Poland with his family in 1918. In 1914, his father Henryk Erlich, a leader in the Socialist movement in Poland, was executed. In the same year, after university studies in Berlin and Warsaw, Erlich emigrated to the United States, where he earned a Ph.D. at the New School for Social Research and joined the faculty of Columbia University in 1955. From 1966 until his retirement in 1981 Erlich was professor of economics at Columbia, teaching in the economics department, the Russian Institute and the Institute for East Central Europe. Professor Erlich was revered by his students for his unstinting help and encouragement and respected by his colleagues for his breadth of knowledge and understanding of socialist economics.

Alexander Erlich's main contribution to the economics of socialism is his work on the critical issue of industrialization policy in the USSR in the 1920s. To this issue, Erlich brought an unusual blend of sophisticated economic reasoning and penetrating political analysis. His major thesis concerning Soviet policy in this period is that the structural disproportions in the Soviet economy were so deep that virtually any policy would have had negative side effects on reconstruction. Specifically, Erlich argued throughout his career that the economic policies of both the left *and* the right opposition were equally problematic. While the left analysis was correct in pointing out that future growth was limited after 1925 by the existing high-capacity utilization and scarce investment funds, Preobrazhenskii and others were wrong in underestimating the reaction of the peasantry to an industrialization policy that would squeeze peasant incomes. On the other hand, the right opposition did not appreciate the implications of high-capacity utilization for continued growth through small profit margins and high turnover of consumer goods and light manufacturers. The right, and Bukharin in particular, were seen by Erlich to be naive on the intensity of the conflict between consumption and investment once existing capacity was fully utilized. This, his major work, exhibits a detailed knowledge of the Soviet experience and a dispassionate and rigorous analysis of policy choices that set the standard for such work in the field.

Selected Works

1950. Preobrazhenskii and the economics of Soviet industrialization. *Quarterly Journal of Economics* 64: 57–88.

1959. The Polish economy after october 1956: Background and outlook. *American Economic Review, Papers and Proceedings* 49: 94–112.

1960. *Soviet industrialization debate, 1924–1928.* Cambridge, MA: Harvard University Press.

1967a. Development strategy and planning: The Soviet experience. In *National economic planning,* ed. M.F. Millikan. New York: Columbia University Press.

1967b. Notes on a Marxian model of capital accumulation. *American Economic Review, Papers and Proceedings* 57: 599–616.

1973. A Hamlet without the Prince of Denmark. *Politics and Society* 4(1): 35–53.

1977. Stalinism and Marxian growth models. In *Stalinism: Essays in historical interpretation,* ed. R.C. Tucker. New York: Norton.

1978. Dobb and the Marx–Fel'dman model: A problem in Soviet economic strategy. *Cambridge Journal of Economics* 2: 203–214.

Bibliography

Desai, P. (ed.). 1983. *Marxism, central planning and the Soviet economy: Economic essays in honour of Alexander Erlich*. Cambridge, MA/London: MIT Press.

Errors in Variables

Vincent J. Geraci

The Historical Ambivalence

This entry surveys the history and recent developments on economic models with errors in variables. These errors may arise from the use of substantive unobservables, such as permanent income, or from ordinary measurement problems in data collection and processing. The point of departure is the classical regression equation with random errors in variables:

$$y = X^*\beta + u$$

where y is a $n \times 1$ vector of observations on the dependent variable, X^* is a $n \times k$ matrix of unobserved (latent) values on the k independent variables, β is a $k \times 1$ vector of unknown coefficients, and u is a $n \times 1$ vector of random disturbances. The matrix of observed values on X^* is

$$X = X^* + V$$

where V is the $n \times k$ matrix of measurement errors. If some variables are measured without error, the appropriate columns of V are zero vectors. In the conventional case the errors are uncorrelated in the limit with the latent values X^* and the disturbances u; and the errors have zero means, constant variances, and zero autocorrelation. In observed variables the model becomes

$$y = X\beta + (u - V\beta).$$

Since the disturbance $(u - V\beta)$ is correlated with X, ordinary least squares estimates of β are biased and inconsistent. The errors thus pose a potentially serious estimation problem. In regard to systematic errors in variables, they will not be discussed, since they raise complex issues of model misspecification which lie outside the scope of this entry.

Errors in variables have a curious history in economics, in that economists have shown an ambivalent attitude toward them despite the universal awareness that economic variables are often measured with error and despite the commitment to economics as a science. Griliches (1974) suggested that much of the ambivalence stems from the separation in economics between data producers and data analysers. If so, why have not economists made a greater effort to cross the breach? Griliches (p. 975) further suggested that 'another good reason for ignoring errors in variables was the absence of any good cure for this disease'. If so, why have not economists made greater use of the econometric techniques developed since Griliches wrote his survey?

We propose an alternative explanation: the way of economic thinking, epitomized by utility theory and consumer maximization, has promoted a neglect of measurement errors. Bentham (1789, ch. IV) was a pioneer of measurement theory in the social sciences in his attempt to provide a theory for the measurement of utility. He went so far as to recommend that the social welfare of a given policy be computed by summing up *numbers* expressive of the 'degrees of good tendency' across individuals. Bentham's notion of cardinal utility met rightfully with great resistance. Pareto (1927, ch. III) pressed the dominant view: the economic equilibrium approach, by producing empirical propositions about consumer demand in terms of observables (quantities, prices, incomes), is to be favoured over theories connecting prices to utility, a metaphysical entity. Thus, theory – here optimization by rational consumers in a competitive market – overcame a fundamental measurement problem.

We do not wish to quarrel with the neoclassical equilibrium approach to the study of demand, although some economists wonder whether the assumptions of the theory have sufficient validity to warrant their acceptance (as part of the

maintained hypothesis) in so many empirical demand studies. Rather, our point is that the great successes of this theory, and analogous successes of similar theories about other economic behaviour, have implanted a subconscious bias toward the substitution of economic theory (assumptions) for difficult measurement. In consequence, many economic models do not have an adequate empirical basis, cf. Leontief (1971) and Koopmans (1979).

Whatever the reasons for their neglect, errors in variables have been hard to keep down. Substantive unobservables such as permanent income, expected price and human capital continue to work their way into economic models and raise measurement issues. Friedman's (1957) permanent income model has served as a prototype for the errors-in-variables setup:

$$c_p = ky_p$$
$$y = y_p + y_t$$
$$c = c_p + c_t$$

where c = consumption, y = income, subscript 'p' = permanent, subscript 't' = transitory, and k is a behavioural parameter. Friedman (p. 36) clearly recognized the connection of the model expressed in observed c and y to the errors-in-variables setup; in his words, 'The estimation problem is the classical one of "mutual regression" or regression "when both variables are subject to error"'.

In the next two sections, early and recent developments on economic models with random errors in variables will be surveyed. Then, we will speculate on the future use of errors-in-variables methods.

Early Econometric Developments

Frisch (1934) was the first econometrician to face squarely the problem of errors in variables. In a brave book addressing model search, multicollinearity, simultaneity, and errors in variables, he decomposed the observed variables into a systematic (latent) part and a random disturbance part. His complicated correlation approach, sometimes resembling common factor analysis, did not satisfactorily resolve the errors-in-variables problem, but he raised fruitful questions. While Koopmans (1937), Geary (1942), Hurwicz and Anderson (1946), Reiersöl (1950), and a few others followed up on Frisch's endeavour, interest in the problem waned by the start of the 1950s. The famous Cowles Commission may have unintentionally buried the errors-in-variables problem when the chief investigators put it aside in order to make progress on the simultaneity problem. Applied economists, in their zeal to employ the new simultaneous equations model, ignored the limitations in their data despite the warning cry of Morgenstern (1950). Sargan (1958), Liviatan (1961), Madansky (1959), and a few others made contributions in the 1950s and 1960s, but for the most part errors in variables lay dormant. Widely used econometrics textbooks aggravated matters by highlighting the lack of identification of the classical regression equation in the absence of strong prior information such as known ratios of error variances. Neglect by the theorists led to the widespread use of *ad hoc* proxies in practice.

Recent Econometric Developments

Zellner (1970) sparked a revival of interest in errors in variables. He attained identification of the permanent income prototype by appending a measurement relation that predicted unobservable permanent income in terms of *multiple causes* (e.g. education, age, housing value), to accompany the natural *indicator* relation in which observed current income is a formal proxy for permanent income. Goldberger (1971, 1972b) stimulated the revival by showing how models with substantive unobservables could be identified and estimated by combining all of the measurement information in a set of multiple equations that arise from multiple indicators or multiple causes. He also drew out the connections among the errors-in-variables model of econometrics, the confirmatory factor analysis model of

psychometrics, and the path analysis model of sociometrics. On the applications side, Griliches and Mason (1972) and Chamberlain and Griliches (1975) studied the important socioeconomic problem of estimating the economic returns to schooling, with allowance for unobservable 'ability'. With Goldberger and Griliches leading the way, the econometric literature on errors in variables flourished in the 1970s.

Multiple Equations

For the permanent income prototype, Zellner's multiple cause relation and indicator relation formed a two-equation measurement system that could be appended to the structural consumption equation. For this three-equation model, Zellner (1970) provided an efficient generalized least squares estimator, and Goldberger (1972a) added a maximum likelihood estimator. This errors-in-variables framework, which can be applied to many situations in which an unobservable appears as an independent variable in an otherwise classical regression equation, has been very useful. Example applications have included Aigner's (1974) study of labour supply in which the wage is an unobservable, Lahiri's (1977) study of the Phillips curve in which price expectations is an unobservable, and Geraci and Prewo's (1977) study of international trade in which transport costs is an unobservable.

Jöreskog and Goldberger (1975) generalized the framework to situations in which there are more than two observed dependent variables. Their model combined prior constraints on the reduced-form coefficients (of the type that arise in econometric simultaneous equations models) with prior constraints on the reduced-form disturbance covariance matrix (of the type that arise in psychometric factor analysis models). For this model which contains multiple indicators and multiple causes (MIMIC) for a single unobservable, they developed a maximum likelihood estimator. Applications of the MIMIC framework have included Kadane et al's (1977) study of the effects of environmental factors on changes in unobservable intelligence over time, and Robins and West's (1977) study of unobservable home value. The framework could be extended to endogenous causes, as Robinson and Ferrara (1977) demonstrated.

Simultaneity

The MIMIC model assumes unidirectional causation. Suppose instead that unobservables appear in a simultaneous equations model. This case calls for less 'outside information' than the single-equation case, since coefficient overidentification, as it would exist in the hypothetical absence of measurement errors, may compensate for the underidentification associated with errors. This idea had appeared in early unpublished works by Hurwicz and Anderson (1946); Goldberger resurrected it in 1971. Geraci (1974), Hausman (1977), and Hsiao (1976) subsequently developed identification analyses and estimators for the simultaneous equations model with errors, taking account of the 'disturbance' covariance restrictions induced by the error structure as well as the usual coefficient restrictions. Many of their results have an instrumental variables interpretation. For illustration, an indicator for an unobservable in a simultaneous equations system is a valid instrumental variable for a given structural equation if the unobservable either (a) does not appear in that equation or (b) appears but has an associated error variance that is identified using information from some other part of the system. Once the latter linkage is recognized, the model well may be identifiable and hence estimable. A notable application has been Griliches and Chamberlain's series of studies on the economic returns to schooling. They employed a triangular model with structural disturbances specified as a function of unobservables (common factors). In various ways their models incorporated simultaneity, multiple indicators, and multiple causes.

Dynamics

Maravall and Aigner (1977), Hsiao (1977), and Hsiao and Robinson (1978) extended the errors-in-variables analysis to dynamic economic models. Among their findings, dynamics are a 'blessing' for identification in that autocorrelation of exogenous variables may provide additional

information; and, upon taking a discrete Fourier transform of the data, many of the results for the contemporaneous model can be carried over to the dynamic model. In the same vein, Geweke (1977) developed a maximum likelihood estimator for the dynamic factor analysis model by reprogramming, in complex arithmetic, Jöreskog's (1970) maximum likelihood algorithm which had developed into the widely used LISREL software package. Geweke applied this estimator to investigate manufacturing sector adjustments to unobservable product demand. Singleton (1977) extended this factor analysis approach to study the cyclical behaviour of the term structure of interest rates. His framework allowed estimation of the model without specifying the causes of the unobservable real rate of interest and price expectations, thus isolating the classic Fisher hypothesis for testing.

As the preceding survey indicates, the recent econometric literature contains many theoretical results on the identification and estimation of structural models that contain substantive unobservables and measurement errors. (For a further survey, see Aigner et al. 1984.) The literature also contains some interesting applications, but not many. Are more forthcoming?

Prospects

We have an uneasy feeling about the state of empirical economics. The development of formal economic theory and associated econometric technique has proceeded at an extraordinary pace. At the same time, what do economists know empirically? Many reported inferences hinge upon model assumptions whose validity remains to be assessed, and the gap between econometric technique and available data seems to be growing. None the less, there are grounds for some optimism.

Although few in number, the applications of errors-in-variables methods in the 1980s have been striking in their relevance to central economic issues. For example, Attfield (1980) has made the permanent income model a more complete explanation of consumption by incorporating unobservable liquid assets, rateable value, and windfall income. His model is a special simultaneous equations model with errors, in which identification of the individual structural equations can be established on a recursive basis. Geweke and Singleton (1981) also have taken up the permanent income model, adapting the classical latent variables model to this time series context and thereby generating some new tests of the permanent income hypothesis. As another example, Garber and Klepper (1980) have defended the competitive model of short-run pricing in concentrated industries through an explicit accounting for errors in measuring cost and output changes. They concluded that short-run price behaviour may appear to be related to market structure primarily because of estimation biases due to the measurement errors. As a final example, Stapleton (1984) has shown that the symmetry restrictions on structural parameters imposed by demand theory can be used to identify a linear model's parameters when measurement errors in price perceptions exist. This study is noteworthy in two respects. First, it shows how price, that bedrock of economic theory, may be measured erroneously. Second, economic theory is used to permit the explicit treatment of errors in variables.

These recent empirical works indicate the potential of errors-in-variables methods to lend fresh insights into important economic issues, and should stimulate more use of these methods. There are other encouraging signs as well. Recent studies using micro data have shown increasing attention to measurement error problems. In the macro area the rational expectations hypothesis has raised economists' consciousness of the difference between key conceptual variables of economic theory (i.e. permanent income, expected price, ex ante real rate of interest) and the available measurements. With respect to applications of errors-in-variables methods in economics, the stock is not great but the flow is encouraging.

See Also

- ▶ Econometrics
- ▶ Latent Variables
- ▶ Regression and Correlation Analysis

References

Aigner, D.J. 1974. An appropriate econometric framework for estimating a labor-supply function from the SEO file. *International Economic Review* 15(1): 59–68.

Aigner, D.J., C. Hsiao, A. Kapteyn, and T. Wansbeek. 1984. Latent variable models in econometrics. In *Handbook of econometrics*, ed. Z. Griliches and M.-D. Intriligator, ch. 23. Amsterdam: Elsevier Science.

Attfield, C.L.G. 1980. Testing the assumptions of the permanent-income model. *Journal of the American Statistical Association* 75: 32–38.

Bentham, J. 1789. *An introduction to the principles of morals and legislation*. London: Clarendon Press. 1907.

Chamberlain, G., and Z. Griliches. 1975. Unobservables with a variance-components structure: Ability, schooling, and the economic success of brothers. *International Economic Review* 16(2): 422–449.

Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.

Frisch, R. 1934. *Statistical confluence analysis by means of complete regression systems*. Oslo: University Institute of Economics.

Garber, S., and S. Klepper. 1980. Administered pricing' or competition coupled with errors of measurement. *International Economic Review* 21(2): 413–435.

Geary, R.C. 1942. Inherent relationships between random variables. *Proceedings of the Royal Irish Academy* 47: 63–76.

Geraci, V.J. 1974. *Simultaneous equation models with measurement error.* PhD dissertation. New York: Garland, 1982.

Geraci, V.J., and W. Prewo. 1977. Bilateral trade and transport costs. *Review of Economics and Statistics* 59(1): 67–74.

Geweke, J.F. 1977. The dynamic factor analysis of economic time-series models. In *Latent variables in socio-economic models*, ed. D.J. Aigner and A.S. Goldberger, ch. 19. Amsterdam: North-Holland.

Geweke, J.F., and K.J. Singleton. 1981. Latent variable models for time series: A frequency domain approach with an application to the permanent income hypothesis. *Journal of Econometrics* 17(3): 287–304.

Goldberger, A.S. 1971. Econometrics and psychometrics. *Psychometrika* 36: 83–107.

Goldberger, A.S. 1972a. Maximum-likelihood estimation of regressions containing unobservable independent variables. *International Economic Review* 13(1): 1–15.

Goldberger, A.S. 1972b. Structural equation methods in the social sciences. *Econometrica* 40(6): 979–1001.

Griliches, Z. 1974. Errors in variables and other unobservables. *Econometrica* 42(6): 971–998.

Griliches, Z., and W.M. Mason. 1972. Education, income and ability. *Journal of Political Economy* 80(3), Pt II, May–June, 74–103.

Hausman, J. 1977. Errors in variables in simultaneous equation models. *Journal of Econometrics* 5(3): 389–401.

Hsiao, C. 1976. Identification and estimation of simultaneous equation models with measurement error. *International Economic Review* 17(2): 319–339.

Hsiao, C. 1977. Identification of a linear dynamic simultaneous error-shock model. *International Economic Review* 18(1): 181–194.

Hsiao, C., and P.M. Robinson. 1978. Efficient estimation of a dynamic error-shock model. *International Economic Review* 19(2): 467–479.

Hurwicz, L., and T.W. Anderson. 1946. Statistical models with disturbances in equations and/or disturbances in variables. Unpublished memoranda. Chicago: Cowles Commission.

Jöreskog, K.G. 1970. A general method for the analysis of covariance structures. *Biometrika* 57(2): 239–51.

Jöreskog, K.G., and A.S. Goldberger. 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 70, Pt I, September, 631–639.

Kadane, J.B., T.W. McGuire, P.R. Sanday, and P. Stuelin. 1977. Estimation of environmental effects on the pattern of I.Q. scores over time. In *Latent variables in socioeconomic models*, ed. D.J. Aigner and A.S. Goldberger, ch. 17. Amsterdam: North-Holland.

Koopmans, T.C. 1937. *Linear regression analysis of economic time series*. Haarlem: De Erven F. Bohn N.V.

Koopmans, T.C. 1979. Economics among the sciences. *American Economic Review* 69(1): 1–13.

Lahiri, K. 1977. A joint study of expectations formation and the shifting Phillips curve. *Journal of Monetary Economics* 3(3): 347–357.

Leontief, W. 1971. Theoretical assumptions and non-observed facts. *American Economic Review* 61(1): 1–7.

Liviatan, N. 1961. Errors in variables and Engel curve analysis. *Econometrica* 29: 336–362.

Madansky, A. 1959. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* 54: 173–205.

Maravall, A., and D.J. Aigner. 1977. Identification of the dynamic shock-error model. In *Latent variable models in socioeconomic models*, ed. D.J. Aigner and A.S. Goldberger, ch. 18. Amsterdam: North-Holland.

Morgenstern, O. 1950. *On the accuracy of economic observations*, 2nd ed. Princeton: Princeton University Press. 1963.

Pareto, V. 1927. *Manual of political economy.* New York: Augustus M. Kelley. 1971.

Reiersl, O. 1950. Identifiability of a linear relation between variables which are subject to error. *Econometrica* 18: 375–389.

Robins, P.K., and R.W. West. 1977. Measurement errors in the estimation of home value. *Journal of the American Statistical Association* 73: 290–294.

Robinson, P.M., and M.C. Ferrara. 1977. The estimation of a model for an unobservable variable with endogenous causes. In *Latent variable models in socioeconomic models*, ed. D.J. Aigner and A.S. Goldberger, ch. 19. Amsterdam: North-Holland.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.
Singleton, K.J. 1977. The cyclical behavior of the term structure of interest rates. Unpublished PhD dissertation. Madison: University of Wisconsin.
Stapleton, D. 1984. Errors-in-variables in demand systems. *Journal of Econometrics* 26(3): 255–270.
Zellner, A. 1970. Estimation of regression relationships containing unobservable independent variables. *International Economic Review* 11: 441–454.

Estate and Inheritance Taxes

Wojciech Kopczuk

Abstract

This article briefly describes features of real-life estate and inheritance taxes, economic arguments for and against these types of taxation and empirical evidence on economic distortions associated with such instruments.

Keywords

Bequests; Capital gains and losses; Charitable contributions; Estate taxation; Gift taxation; Inheritance taxation; Progressive and regressive taxation; Redistribution of income and wealth; Tax avoidance; Taxation of income; Wealth accumulation

JEL Classifications

H2

Taxes imposed on intergenerational transfers are among the oldest types of taxation, apparently dating back at least to the Roman Empire (Pechman, 1987). There is substantial variation in their design in actual tax systems. The tax may be imposed on the donor or the donee side: it can apply either to the total estate (the total value of assets left by the decedent) or it can apply separately to transfers received by each beneficiary. This distinction matters when there are multiple beneficiaries and the tax is not simply proportional, or if the inheritance tax interacts with other forms of taxation (such as income tax). Further sources of variation in how these types of tax appear around the world include differences in how family members are treated, deductions allowed, treatment of certain categories of assets, treatment of capital gains and interaction with other types of tax. Two additional types of tax are closely associated with estate and inheritance taxation. First, some countries impose additional tax on transfers that skip generations. Such transfers would otherwise avoid taxation at death of an intermediate generation and hence would provide tax savings. Second, taxes on *inter vivo* gifts are imposed to protect the base of estate taxation (this is not their sole purpose, however: they also reduce the incentive for income shifting across individuals subject to different individual income tax brackets).

Most developed countries impose some form of taxation of intergenerational transfers; the exceptions are Canada, Australia and New Zealand. European countries usually impose inheritance taxes. See Gale and Slemrod (2001) for more details.

Estate Taxation in the United States

In the United States, estate and gift taxes are 'integrated', that is, gifts over the lifetime influence computation of the estate tax burden at death. On top of federal taxation, many states impose their own taxes (in some cases inheritance rather than estate), although since the 1970s most states have changed their taxes to only 'soak up' federal credit for state taxation without imposing any incremental tax liability for those who are subject to the federal tax. The modern federal estate tax was introduced in 1916, although many states imposed their own taxes before that and the federal government made two earlier attempts to tax estates (during the Civil War and the Spanish–American War). The structure of estate taxation changed often before the Second World War, when the top marginal tax rates hit 77 per cent. Marginal tax rates were not reduced until the early 1980s, when the top rate was cut to 55 per

cent. Further reductions are taking place as a part of the phase-out of estate tax initiated in 2001 that is supposed to culminate in a repeal in 2010. The repeal is a part of the set of provision that sunset in 2011, and hence the future of this tax is uncertain at the time of this writing. The US estate tax has always been characterized by a large tax exemption. At the peak in 1976, slightly over seven per cent of adult deaths corresponded to taxable estates, but, other than during the period of the growth in the reach of the tax in the 1960s and 1970s induced by 'bracket creep' (brackets not indexed for inflation), only two per cent or less of all estates were subject to the tax. Revenue collected by this tax has always been relatively small, constituting one to two per cent of total federal revenue after the Second World War.

Arguments for and Against Estate and Inheritance Taxation

A number of arguments are often given in favour of this type of taxation:

- Administrative convenience – taxation occurs at the time when assets have to be valued anyway, thereby reducing the burden of compliance relative to other forms of wealth taxation.
- Presumed lack of distortions if bequests are mostly 'accidental', that is, when taxpayers save for their own lifetime consumption rather than for bequests (see Kopczuk, 2003, for a critique of this argument).
- Redistribution (although, Kaplow, 2001 suggests that income taxation may be sufficient for redistribution).
- Backstop to avoidance of income taxes.
- Providing equality of opportunities and breaking down concentration of wealth.
- 'Carnegie effect' – inherited wealth is a 'bad' because it makes children unproductive members of society (see Holtz-Eakin, Joulfaian and Rosen, 1993, for supporting empirical evidence).
- Providing incentives for charity.

These are countered by the following:

- Distortions introduced by estate taxation.
- Theoretical arguments for zero capital taxation in the long run (recently challenged in the context of the estate tax by Farhi and Werning, 2007).
- Horizontal inequity due to unequal treatment of 'savers' and 'spenders' (see for example McCaffery, 1994).
- Easy tax avoidance.
- Gift externality (providing an argument for subsidizing transfers).

A broader overview of the normative issues can be found in Gale and Slemrod (2001) and Kaplow (2001).

Economic Distortions

Among the types of economic distortions often discussed in this context are effects on saving, investment and labour supply, tax avoidance and damage that is potentially done to small (family) firms when the owner dies (see Brunetti, 2006, for weakly supporting evidence that survival of small businesses is affected by the presence of this tax; note, though, that small firms already enjoy significant preferences in the US tax code). A related argument involves forcing taxpayers to pursue 'deathbed' planning and implications of the tax for 'widows and orphans' (see Kopczuk, 2007). Because of the presence of a deduction for charitable contributions in the United States, an important topic is the effect of the estate tax on charitable contributions. Some of the more important empirical findings regarding US estate tax are discussed in what follows.

Estate tax avoidance is thought to be very easy. Cooper (1979) suggested that a motivated tax planner could easily reduce tax liability very significantly, if not altogether. Others have challenged this view: for example, Schmalbeck (2001) argues that most avoidance strategies involve losing control over assets. There is some evidence in support of both views. Anecdotal evidence of widespread estate tax avoidance is

easy to obtain, and the existence of a large estate tax planning industry is a prima facie evidence that a lot of effort goes into such planning. At the same time, it has been established that some simple tax avoidance strategies are not pursued enough from the tax minimization point of view. McGarry (1999) and Poterba (2001) show that taxpayers do not take full advantage of annual gift tax exemption (annual gifts of less than $11,000 per donee are exempt from taxation). Kopczuk (2007) shows that significant adjustments take place following the onset of a terminal illness, thereby revealing that not enough planning took place earlier in life. Kopczuk and Slemrod (2003) argue that the widespread reliance by married decedents on the unlimited marital deduction implies that taxpayers do not take full advantage of tax savings from splitting an estate. This is so despite the existence of trust instruments that allow for separating tax planning from other considerations such as taking care of the surviving spouse. On the other hand, while gifts do not seem to be fully utilized as a tax-planning device, they are nevertheless responsive to tax considerations, as demonstrated by Bernheim, Lemke and Scholz (2004) and Joulfaian (2004).

A number of papers have focused on estimating the responsiveness of estates to tax rates. Kopczuk and Slemrod (2001), Holtz-Eakin and Marples (2001) and Joulfaian (2006) all found small but positive elasticities implying that higher marginal tax rates lead to lower estate values. Kopczuk and Slemrod (2001) and Joulfaian (2006) rely on estate tax data and therefore cannot distinguish between tax avoidance and the effect on wealth accumulation. Holtz-Eakin and Marples (2001) use actual wealth, but their results are based on a relatively low-wealth sample and hence are hard to generalize from. Due to the nature of estate taxation, these studies are based on cross-section, repeated cross-section or time series, and hence the econometric assumptions that underlie them are strong.

The estate tax is a part of the tax code, and considering it in isolation is not appropriate. Poterba and Weisbenner (2001) find that over 50 per cent of the value of estates over $10 million are unrealized capital gains that would escape taxation at death due to step-up provisions (capital gains unrealized at the time of death are not subject to the capital gains tax, and the base for the recipient is stepped up to the current value of the asset). Auten and Joulfaian (2001) find that lower estate tax rates reduce capital gains realizations, and thus exacerbate the lock-in effect. The estate tax constitutes a backstop to this type of avoidance and a repeal of the tax would require a modification of the step-up rule. Bernheim (1987) questioned whether the estate tax raises any net revenue once its interaction with other taxes is taken into account.

The effect of estate taxes on charitable contributions is theoretically ambiguous due to offsetting income and substitution effects. Most studies find that higher marginal estate tax rates stimulate charitable giving, but the magnitude of the overall effect, accounting for both price and wealth effects, remains controversial: Joulfaian (2005) provides a recent overview of the empirical literature.

Other than dealing with tax avoidance (the issue that may be better handled by fixing the income tax), the strongest arguments in favour of the tax are based on its role in redistribution, breaking up concentration of wealth and providing equality of opportunities. Kopczuk and Saez (2004) use historical estate tax return data to provide estimates of wealth concentration over the course of the 20th century, and discuss the role that the estate tax might have played in shaping trends in concentration. Piketty and Saez (2007) document the contribution of the estate tax to overall progressivity. Understanding how estate taxation influences the distribution of wealth should be the top priority for anyone interested in an honest assessment of its value as a policy instrument.

See Also

- ▶ Bequests and the Life Cycle Model
- ▶ Capital Gains Taxation
- ▶ Excess Burden of Taxation
- ▶ Inheritance and Bequests

- ▶ Optimal Taxation
- ▶ Redistribution of Income and Wealth
- ▶ Tax Compliance and Tax Evasion
- ▶ Taxation of Income
- ▶ Taxation of Wealth

Bibliography

Auten, G., and D. Joulfaian. 2001. Bequest taxes and capital gains realizations. *Journal of Public Economics* 81: 213–229.

Bernheim, B.D. 1987. Does the estate tax raise revenue? In *Tax policy and the economy*, vol. 1, ed. L.H. Summers. Chicago: NBER, Cambridge, MA: MIT Press.

Bernheim, B.D., R. Lemke, and J.K. Scholz. 2004. Do estate and gift taxes affect the timing of private transfers? *Journal of Public Economics* 88: 2617–2634.

Brunetti, M. 2006. The estate tax and the demise of the family business. *Journal of Public Economics* 90: 1975–1993.

Cooper, G. 1979. *A voluntary tax? New perspectives on sophisticated tax avoidance*. Washington, DC: Brookings Institution Press.

Farhi, E., and I. Werning. 2007. Progressive estate taxation. Mimeo: MIT Department of Economics. http://econ-www.mit.edu/faculty/download_pdf.php?id=1403.

Gale, W.G. and J. Slemrod 2001. Rethinking the estate and gift taxes: Overview. In eds. Gale, Hines and Slemrod.

Gale, W.G., J.R. Hines Jr., and J. Slemrod, eds. 2001. *Rethinking estate and gift taxation*. Washington, DC: Brookings Institution Press.

Holtz-Eakin, D., and D. Marples. 2001. Distortion costs of taxing wealth accumulation: Income versus estate taxes. Working Paper No. 8261. Cambridge, MA: NBER.

Holtz-Eakin, D., D. Joulfaian, and H.S. Rosen. 1993. The Carnegie conjecture: Some empirical evidence. *Quarterly Journal of Economics* 108: 413–435.

Joulfaian, D. 2004. Gift taxes and lifetime transfers: Time series evidence. *Journal of Public Economics* 88: 1917–1929.

Joulfaian, D. 2005. Estate taxes and charitable bequests: Evidence from two tax regimes. Working Paper No. 92. Office of Tax Policy Analysis, US Department of the Treasury.

Joulfaian, D. 2006. The behavioral response of wealth accumulation to estate taxation: Time series evidence. *National Tax Journal* 59: 253–268.

Kaplow, L. 2001. A framework for assessing estate and gift taxation. In eds. Gale, Hines and Slemrod.

Kopczuk, W. 2003. The trick is to live: is the estate tax social security for the rich? *Journal of Political Economy* 111: 1318–1341.

Kopczuk, W. 2007. Bequest and tax planning: Evidence from estate tax returns. *Quarterly Journal of Economics* 122(4).

Kopczuk, W., and E. Saez. 2004. Top wealth shares in the United States: Evidence from estate tax returns. *National Tax Journal* 57: 445–488.

Kopczuk, W. and Slemrod, J. 2001. The impact of the estate tax on the wealth accumulation and avoidance behavior of donors. In Gale, Hines and Slemrod (2001).

Kopczuk, W., and J. Slemrod. 2003. Tax consequences on wealth accumulation and transfers of the rich. In *Death and dollars: The role of gifts and bequests in America*, ed. A.H. Munnell and A. Sunden. Washingon, DC: Brookings Institution Press.

McCaffery, E. 1994. The uneasy case for wealth transfer taxation. *Yale Law Journal* 104: 283–365.

McGarry, K. 1999. Inter vivos transfers and intended bequests. *Journal of Public Economics* 73: 321–351.

Pechman, J. 1987. Inheritance taxes. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, Vol. 2. London: Macmillan.

Piketty, T., and E. Saez. 2007. How progressive is the U.S. federal tax system? An international and historical perspective. *Journal of Economic Perspectives* 21(1): 3–24.

Poterba, J.M. 2001. Estate and gift taxes and incentives for inter vivos giving in the US. *Journal of Public Economics* 79: 237–264.

Poterba, J.M. and S. Weisbenner 2001. The distributional burden of taxing estates and unrealized capital gains at death. In eds. Gale, Hines and Slemrod.

Schmalbeck, R. 2001. Avoiding federal wealth transfer taxes. In eds. Gale, Hines and Slemrod.

E

Estimation

Marc Nerlove and Francis X. Diebold

Point estimation concerns making inferences about a quantity that is unknown but about which some information is available, e.g., a fixed quantity θ for which we have n imperfect measurements $x_1,\ldots,x_n$) The theory of estimation deals with how best to use the information (combine the values $x_1,\ldots,x_n$) to obtain a single number, estimate, for θ, say $\widehat{\theta}$. Interval estimation does not reduce the available information to a single number and is a special case of hypothesis testing. This entry deals only with point estimation.

Justification for any particular way of combining the available information can be given only in

terms of a *model* connecting the x's to θ. For example, in the case of imperfect measurements $x_1,\ldots, x_n$, we could regard the *errors,* $x_i - \theta$, $i = 1,\ldots, n$ as independent outcomes of a random process so that the joint distribution of the x's depends on θ:

$$p(x_1, \ldots, x_n|\theta) = \prod_1^n f(x_i - \theta).$$

In general, a *statistical model* represents the data, observations $x_1,\ldots,x_n$, x, where the x's may be vectors of quantities, as having arisen as a drawing from a joint distribution depending on some unknown parameters $\theta = (\theta_1,\ldots, \theta_k)'$. For example, consider $x_1,\ldots,x_t$, where x_t, is identically and independently distributed according to a univariate normal distribution with mean μ and variance σ^2 (Cramer 1946). The "location parameter," μ, and the "scale parameter, " σ^2, are unknown but, because they determine the distribution from which the data are supposed to arise, the latter may be used to form a point estimate of the vector $\theta = (\mu,\sigma^2)'$, e.g., $\widehat{\theta} = \left(\overline{x}\Sigma_1^T x_i/T, s^2 = \Sigma_1^T(x_i - \overline{x})^2/T\right)'$, the properties of which may be discussed in terms of various criteria and the properties of the family of probability distributions $p(x|\ \theta)$ from which the data are assumed to come. An *estimator* is a *function* of the observations; an *estimate*is the value of such a function for a particular set of observations. The theory of point estimation concerns the justification for estimators in terms of the properties of the estimates which they yield relative to specified criteria.

General treatments of the theory of point estimation may be found in Lehmann (1983), Cox and Hinkley (1974), Rao (1973) and Zellner (1971), *inter alia*.

Econometric estimation problems usually concern inferences about the parameters of conditional rather than unconditional distributions. For example, if the observations $(y_1, x_1,\ldots, (y_n, x_n)$, are assumed to represent a drawing from a multivariate normal distribution with mean vector μ and variance-covariance matrix Σ , then the *conditional distribution of y* given *x,* $p(y \mid x, \theta)$, is univaraite normal with mean $\theta_1 = \mu_1 + \sigma_{12}\sigma_{22}^{-1}(x - \mu_2)$ and variance $\theta_2 = \sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma_{21}$, where

$$\mu = (\mu_1, \mu_2) \text{ and } \sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

Note that θ_1 is a linear function of x which depends upon the parameters of the originally assumed joint distribution; this function is called the *regression* of y on *x. regression analysis* deals with the general problem of estimating such functions which characterize conditional distributions, usually those derived from normal distributions.

A standard method, and the one most common in econometrics, for obtaining estimators is the method of *maximum likelihood*. Consideration of this method provides a good introduction to alternative principles of estimation. Let the data $x = (x_1,\ldots, x_n)'$ be fixed and regard $p(x \mid \theta)$ as a function of θ it is then called the *likelihood*. The value of $\widehat{\theta} = \widehat{\theta}\,(x_1,\ldots,x_n)$ which maximizes $p(x \mid \theta)$, if it is exists and is unique, is called the maximum-likelihood estimator, or estimate (MLE). (For a general survey, see Norden 1972–1973, or Lehmann 1983.) The MLE of a continuous function $g(\theta)$ is $g(\widehat{\theta})$ *where* $\widehat{\theta}$ is the MLE of θ. Other desirable properties of the MLE are asymptotic as $n \to \infty$. Under regularity conditions: (1) The MLE is weakly *consistent*, i.e., $\lim n \to \infty \Pr\left(|\ \widehat{\theta}_n - \theta\right| < \in) = 1$ for all $\in > 0$. (2) The MLE is *asymptotically normal*, i.e. the distribution of $\widehat{\theta}$ appropriately normalized, $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$, tends to the normal distribution, with mean O and variance-covariance matrix $[I(\theta)]^{-1}$ where

$$I(\theta) = -E\left[\partial^2 \log p(x|\theta)/\partial\theta\partial\theta'\right].$$

I(θ) is called the information matrix and shows the information a single observation contains about the parameter θ. (3) The MLE is *asymptotically efficient* in the sense that if θ^* is any other estimator such that $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$ tends in distribution to the normal with mean zero and variance-covariance matrix $\sum(\theta)$, the matrix $[\sum(\theta) - I^{-1}(\theta)]$ is positive semi-definite. For example, in the case of one parameter this means

that no other asymptotically normal estimator has, as $n \to \infty$, a smaller variance than the MLE. The conditions for asymptotic normality do ensure, with probability tending to one, a solution to the *likelihood equation* $\partial \log p(x|\theta)/\partial\theta = 0$, which is consistent and asymptotically normal and efficient. The problem is that there may be more than one solution, but only one can be the MLE. When the number of parameters to be estimated (elements of the vector θ) tends to infinity with n, the MLE's for some may exist but may not be consistent (Neyman and Scott 1948).

Solutions to the likelihood equation are not the only estimators which may be consistent, asymptotically normal and efficient, but comparison with the MLE, assuming correct specification of $p(x|\theta)$, is facilitated by the fact that all have a normal distribution as $n \to \infty$. For fixed n, the distributions of different estimators are difficult to determine and may, indeed, be quite different. Moreover, when the distributions underlying the data are misspecified, the MLE's generally no longer have these optimal properties (White 1982; Gourieroux et al. 1984), although other, weaker, optimality properties remain. Apart from specification problems, however, the likelihood function provides an important and useful summary of the data, and point estimates and hypothesis testing procedures based on it are often justified in this way (Fisher 1925; Barnard et al. 1962; Edwards 1972).

The 'accuracy' of an estimator $\widehat{\theta}$ of a scalar parameter θ may be measured (defined) in a variety of ways: by its expected squared or absolute error, relative error, or by $\Pr\{|\widehat{\theta} - \theta| \leq \alpha\}$ for some α. Any choice is arbitrary; for convenience expected squared error is the usual choice. Some justification for a particular choice may be provided in terms of a *loss function* $L(\theta, \widehat{\theta})$ *or the expected loss* $EL(\theta, \widehat{\theta})$ *or risk function* of *statistical decision theory*. Choice of estimators may be justified in terms of the extent to which the choice minimizes risk or some aspect thereof. Both the *sampling theory* and *Bayesian* approaches to estimation can be interpreted in these terms.

A very weak property that any estimator should have is that no other estimator exists which dominates it in the sense that the latter leads to estimates having uniformly lower expected loss irrespective of θ. Estimators satisfying this criterion are called *admissible*.

In the sampling theoretic approach, emphasis is placed on finding estimators which have desirable properties in terms of relative frequencies in hypothetically repeated samples. For example, we might require that the distribution of an estimator be centred on the true parameter value, i.e., $E(\widehat{\theta} - \theta) = 0$. *Such estimators are called unbiased*. Among all unbiased estimators we presumably would prefer one yielding estimates with a distribution concentrated about the mean. Such minimum variance unbiased estimators (MVU) play a key role in the theory of estimation. Specifically, the famous Rao-Blackwell Theorem states that if an unbiased estimator $\widehat{\theta}$ is a function of a complete sufficient statistic for θ then it is MVU. A statistic, say T, is said to be sufficient for θ if the conditional distribution of the observations given T is independent of θ. Completeness is also a property of the distribution functions for the observations; (a family P of distributions (of T) indexed by a parameter θ is said to be complete if there is no 'unbiased estimator of zero' other than $\Phi(x) \equiv 0$.) Note that choosing an estimator so as to minimize the expected squared error of the estimate it yields is equivalent to minimizing the unweighted sum of the variance and the squared bias. From a decision theoretic point of view, it may be better to accept an estimator with a small bias if such an estimator has a smaller risk.

In the sampling theoretic approach, emphasis is given to the distribution of estimates yielded by a specified estimator. The likelihood approach, on the other hand, emphasizes the distribution of the observations, given a parametrically specified distribution, under alternative values of these parameters. Concern is primarily with the maximum value of the likelihood function with respect to the parameters and its curvature near the point at which the global maximum occurs, but some approaches stress the relevance of the likelihood function in other neighbourhoods (Barnard et al. 1962; Edwards 1972). The Bayesian approach carries concern with the entire

likelihood function further: estimation and inference are based on the posterior density of the unknown parameters of the distribution generating the observations. This posterior density is proportional to the likelihood function multiplied by a prior density of the parameters, i.e., a weighted average of likelihoods for different parameter values where the weights are determined by prior (subjective) beliefs. (See BAYESIAN INFERENCE.)

In the Bayesian approach, both observations and parameters are taken to be stochastic. Let $p(x_1, \theta)$ be the joint probability density function for an observation vector, x, and a parameter vector θ; then $p(x_1, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$, where $p(\xi|\eta)$ denotes the conditional density of ξ given η and $p(\xi)$ denotes the marginal density of ξ. Thus $p(\theta|x)$ *is proportional to* $p(\theta)$ $p(x|\theta)$ by the factor

$$p(x) = \int p(\theta)p(x|\theta)\mathrm{d}\theta.$$

$p(\theta|x)$ is the *posterior* distribution of $^{\theta}$ after having observed the data; $p(\theta)$ is the *prior* distribution of θ and $p(x|\theta)$ is the *likelihood*. Alternatively, consider the weighted average risk (as defined above):

$$\int EL\left(\theta, \widehat{\theta}\right) w(\theta)\mathrm{d}\theta,$$

with weights $w(\theta)$ such that

$$\int w(\theta)\mathrm{d}\theta = 1.$$

When $L(\theta, \widehat{\theta}) = (\widehat{\theta} - \theta)^2$, the estimator which minimizes such a weighted average risk is

$$\widehat{\theta}(x) = \int \theta w(\theta)p\left(\theta w(\theta)p(x|\theta)\right. \mathrm{d}\theta \int w(\theta)p(x|\theta) \times \mathrm{d}\theta.$$

If the *weights* $w(\theta)$ are taken to be the values of the marginal density $p(\theta)$, the mean of the posterior Bayes distribution minimizes the expected squared error of the estimates when both the variation of data and the uncertainty with respect to θ are taken into account: $\widehat{\theta}$ *posterior* distribution of θ.

As is the expected value of θ based on the $n \to \infty$, it may be shown that the influence of the prior distribution diminishes until in the limit it disappears; then, under general circumstances, the minimization of mean square error in the Bayesian framework yields the MLE. The principal difficulty in the Bayesian approach is the choice of a reasonable prior for θ, $p(\theta)$. (For a comprehensive discussion, see Zellner 1971.)

Instead of minimizing the expected loss, one may minimize the maximum loss. Estimators which do are called *minimax*; the theory is developed in Wald (1950).

There are three general approaches to choice of a prior in Bayesian analysis. First, the prior may be obtained empirically (Maritz 1970). For example, suppose that the problem is to estimate the percentage of defective items in a particular batch. Assuming such batches were produced in the past suggests a prior based on the proportion of defective items observed in previous batches. This kind of 'updating' forms the basis for the celebrated Kalman filter. Second, the prior may be viewed as representing a 'rational degree of belief' (Jeffreys 1961). What represents a 'rational degree' is not specified, but the idea leads directly to the use of priors that represent knowing little or nothing, so-called *non-informative priors*. However, total ignorance has proved difficult to capture in many cases. A third approach is that the prior represents a subjective degree of belief (Savage 1954; Raiffa and Schlaifer 1961). But, of whom? and how arrived at? Minimax-estimation theory offers one possible approach for it leads to the minimum mean-square-error Bayes estimator, i.e., the mean of the posterior distribution of the parameters, when the prior is least favourable in the sense of making expected loss the largest for whatever class of priors is chosen.

Related to this problem is the more general question of *robust estimation*. In order to make sense of any data, it is necessary to assume something. For example, the justification for using the sample mean to estimate the mean of the distribution generating the data is often the assumption that that distribution is normal or nearly so. In that

case, the sample mean is not only asymptotically efficient but uniformly MVU, minimax, admissible, etc. But suppose that the distribution is Cauchy (having roughly the same shape as the normal but with very thick tails); then, the sample mean has the same distribution as any individual observation, its accuracy does not improve with n and it is not even a consistent estimator. At least, within the class of distributions which include the Cauchy, the properties of the sample mean, and similarly ordinary least squares, are quite sensitive to the true nature of the underlying distribution of the data. We say that such estimators are not *robust*. Complete discussions are contained in Huber (1981) and Hampel et al. (1985).

To conclude, three estimation problems of special concern in economics are discussed: (1) classical linear regression; (2) non-linear regression, and (3) estimation of simultaneous structural equations.

The classical theory of linear regression deals with the following problem: Let X be an $n\ k$ matrix of nonstochastic observations (n for each variable $(x_1,\ldots, x_k)$, β be a k 1 vector of parameters (one of which becomes an intercept if $x_1 \equiv 1$, say), and y be an n 1 vector of stochastic variables such that $y - x\beta = \in \sim N\ (0,\ \Sigma)$. The *ordinary least-squares estimates* (OLS), $\widehat{\beta} = (X'\ X)^{-1}X'y$ are MLE and MVU when $\Sigma = \sigma^2$ When this is not true, although the OLS estimates are unbiased and consistent, they are not asymptotically efficient or minimum variance. The *generalized least squares estimates* (GLS), $\widehat{\beta}\,(X'\ \Sigma^{-1}\ X)^{-1}X'\ \Sigma^{-1}$ are efficient, but of course Σ and therefore $\Sigma^{-1,}$ is generally unknown. Often, however, a consistent estimate of Σ is available, leading to *feasible*, or *estimated*, GLS estimates.

Many problems in economics lead to non-linear relationships. Linear regression may be a good (local) approximation to such relationships if the data do not vary too widely. Moreover, many non-linear relationships may be transformed into linear ones (e.g., the Cobb-Douglas production function). Often, however, the data are sufficiently variable to make a linear relationship a poor approximation and no linearizing transformation exists. The general non-linear regression model is $y = f(X,\ \beta,\ \in)$ or more frequently $y = f(X,\ \ \beta) + \in$. Leastsquares or maximum-likelihood estimates may still be obtained, but the first-order conditions for a minimum or a maximum will generally be non-linear, frequently ruling out analytic expressions for the estimates. Consider the problem of minimizing the sum of squared residuals, $(y - f(X,\ \beta))'$ $(y - f(x,\beta))'$ with respect to β (non-linear least squares); numerical methods for solving this problem are of the general form: $\widehat{\beta}_{i+1} = \widehat{\beta}_i - s_i P_i$ ∇_i, where $\widehat{\beta}_i$ = the value of the estimator parameter vector at iteration i, s_i = the step size at iteration i, P_i = the direction matrix at iteration i, and ∇_i = the gradient of the objective function at iteration i. The matrix P_i, determines the direction in which the parameter vector is changed at each iteration; it is generally taken to be the Hessian matrix evaluated at the current value of the parameter vector or some approximation to it. Let $g(\beta)$ be the objective function; then

$$P_i = \left[\partial^2 g(\beta)/\partial\beta\partial\beta\prime \middle| \beta = \widehat{\beta}_i\right]^{-1}$$

is the Hessian. A justification for this choice is obtained from the second-order (quadratic) approximation to the objective function in the neighborhood of the current estimate. For a detailed treatment of this problem as well as constrained non-linear estimation see Quandt (1983). The statistical properties of non-linear estimators are discussed by Amemiya (1983).

Economic theory teaches us that the values of many economic variables are often determined simultaneously by the joint operation of several economic relationships, for example, supply and demand determine price and quantity. This leads to a representation in terms of a system of simultaneous structural equations (simultaneous equations model, or SEM). The problem of how to estimate the parameters of an SEM has occupied a central place in econometrics since Haavelmo (1944). A linear SEM is given by, $By_i\ \Gamma x_t = \mathrm{u}_{\mathrm{t}}$, $t = 1,\ldots, T$ where B is $G \times G$, Γ is $G \times K$, y_t is $G \times 1$, x_t is $K \times 1$, and u_t is $G \times 1$. u_t is assumed to be zero mean with variance-covariance matrix Σ often normally distributed, independently and

identically for each t. Thus the u_t are serially independent. It is also assumed that plim $\Sigma_1^T x_{it}u_{jt}/T = 0$ all $i = 1,\ldots, K$ and $j = 1,\ldots,G$ and plim $X'X|T$ is a positive definite matrix, where $X = (x_1,\ldots, x_T)'$. If B is non-singular this system of *structural equations*, as they are called, may be solved for the so called 'endogenous' variables, y_t, in terms of the 'exogenous' variables x_t: $y = \Pi x_t + v_t$ *where* $\Pi = -B^{-1}\Gamma$, $v_t = B^{-1}u_t$, so that $Ev_t = 0$;and $Ev_t v'_t = B^{-1}\ \Sigma\ (B^{-1})' = \Omega$. *It is, in general, not possible to determine* B, Γ and Σ from knowledge of the *reduced form* (*RF*) parameters Π and Ω there are, in principle, many structural systems compatible with the same RF. Given sufficient restrictions on the structural system, however, knowledge of the RF parameters can be used, together with the assumed restrictions, to determine the structural parameters. The SEM is then said to be *identified*.

For *linear* structural equations with *normally distributed* disturbances, the conditions for* identification may be derived from the condition that for any system $B^*y_t + \Gamma^*x_t = u_t^*$ *for which* u_t and u_t, are identically distributed, where $B^* = FB$, $\Gamma^* = F\,\Gamma$and $u_t^* = Fu_t$, then $F \equiv fI$ is implied by the restrictions, where f is any positive scalar (Hsiao 1983).

Methods of estimating the parameters of SEMs may be put into two categories: (1) *limited-information* methods which estimate parameters of a subset of the equations, usually a subset consisting of a single equation, taking into account only the identifying restrictions on the parameters of equations in that subset, and (2) *full-information* methods which estimate all of the identifiable parameters in the system simultaneously and therefore take into account all identifying restrictions. Full- or limited-information methods may be based on either least-squares or maximum-likelihood principles. ML-based methods yield estimates which are invariant according to the normalization rule (choice of f).

For systems or single equations in SEMs for which there are restrictions just sufficient to identify the parameters of interest, estimates may be based on *indirect least squares*, that is, derived directly from the reduced form parameters estimated by applying OLS to each equation of the RF; such estimates are ML. If the restrictions are just sufficient to identify the parameters of each equation, the resulting estimates are *full-information maximum-likelihood* (FIML) estimates. When an equation is over-identified, in the sense that there are more than enough restrictions to identify it, *two-stage least squares* (2SLS) or *limited-information maximum likelihood* (LIML) may be applied equation by equation to each equation which is identified. Provided the model is correctly specified, such estimates are consistent and asymptotically unbiased but not asymptotically efficient, because some restrictions are neglected in the estimation of some parameters. An analog of 2SLS, *three-stage least squares* (3SLS), yields estimates which are asymptotically equivalent to FIML and therefore efficient.

Amemiya (1983) extends all of these methods to non-linear systems. Sargan (1980) discusses identification in non-linear systems.

See Also

- Bayesian Inference
- Least Squares
- Likelihood
- Regression and Correlation Analysis
- Residuals
- Statistical Decision Theory
- Statistical Inference

Bibliography

Amemiya, T. 1983. Nonlinear regression models. In *Handbook of econometrics*, vol. 1, ed. Z. Griliches and M.-D. Intriligator. Amsterdam: North-Holland.

Barnard, G.A., G.M. Jenkins, and C.B. Winsten. 1962. Likelihood inference and time series (with discussion). *Journal of the Royal Statistical Society, Series A* 125: 321–372.

Cox, D.R., and D.V. Hinkley. 1974. *Theoretical statistics*. London: Chapman & Hall.

Cramer, H. 1946. *Mathematical methods of statistics*. Princeton: Princeton University Press.

Edwards, A.W.F. 1972. *Likelihood*. Cambridge: Cambridge University Press.

Fisher, R.J. 1925. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22: 700–725.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica* 52: 681–700.
Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement): 1–115.
Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. 1985. *Robust statistics*. New York: Wiley.
Hsiao, C. 1983. Identification. In *Handbook of econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator, 223–283. Amsterdam: North-Holland.
Huber, P.J. 1981. *Robust statistics*. New York: Wiley.
Jeffreys, H. 1961. *Theory of probability*, 3rd ed. Oxford: Clarendon Press.
Lehmann, E.L. 1983. *The theory of point estimation*. New York: Wiley.
Maritz, J.S. 1970. *Empirical bayes analysis*. London: Methuen.
Neyman, J., and E.L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16: 1–32.
Norden, R.H. 1972–1973. A survey of maximum-likelihood estimation. *Review of the International Institute of Statistics* 40: 329–354; 41: 39–58.
Quandt, R.E. 1983. Computational problems and methods. In *Handbook of econometrics*, vol. 1, ed. Z. Griliches and M. Intriligator, 699–764. Amsterdam: North-Holland.
Raiffa, H., and R. Schlaifer. 1961. *Applied statistical decision theory*. Boston: Harvard Business School.
Rao, C.R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: Wiley.
Sargan, J.D. 1980. Identification and lack of identification. Paper presented to 4th World Congress of the Econometric Society, 28 August–2 September 1980, Aix-en-Provence.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Wald, A. 1950. *Statistical decision functions*. New York: Wiley.
White, H. 1982. Maximum-likelihood estimation of misspecified models. *Econometrica* 50: 1–25.
Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.

Ethics and Economics

Marc Fleurbaey

Abstract

In recent decades, an important corpus of theories in normative economics (social choice theory, the theory of fair allocation, and inequality and poverty measurement in particular) has developed in which formal analytical tools of economic theory are mobilized in order to relate basic principles of social ethics to precise criteria for the evaluation of social states of affairs. The efficacy of arguments based on veil-of-ignorance devices has been questioned and the scope of impossibility theorems has been circumscribed, leaving the stage to a variety of constructive proposals in several fields of application (voting, resource allocation, public policy, social indicators).

Keywords

Arrow, K.; Axiomatic bargaining; Axiomatic theory of indices; Bargaining; Bergson–Samuelson social welfare function; Borda rule; Business ethics; Capability sets; Coalition formation; Competitive equilibrium; Consequentialism; Consumer sovereignty; Cooperative games; Corporate social responsibility; Cost–benefit analysis; Cost–surplus sharing; Dominance criteria; Dworkin, R.; Egalitarian equivalence; Egalitarianism; Equality of capabilities; Equality of opportunity; Equality of resources; Ethics and economics; Fair allocation; Gini coefficient; Harsanyi, J.; Hypothetical insurance (Dworkin); Identity; Impartial observer (J. Harsanyi); Impartiality; Impossibility theorem; Income distribution; Indexing problem; Inequality (measurement); Inequality aversion; Interest aggregation; Interpersonal comparisons of utility; Intertemporal preferences; Judgements vs interests; Liberal paradoxes; Libertarianism; Lorenz curve; Marginal rate of substitution; Maximin principle; Multidimensional inequality; Nash, J.; New Welfare Economics; Normative economics; Objective advantage; Ordinal non-comparable (ONC) preferences; Original position (Rawls); Other-regarding preferences; Pareto principle; Positive economics; Potential Pareto improvements; Poverty measurement; Preferences; Rawls, J.; Responsibility; Rights; Risk aversion; Samuelson, P.; Sen, A.; Shapley, L.; Social choice; Social contract; Social justice; Social welfare functions; Subjective utility;

Tastes; Utility functions; Utility possibility sets; Value judgements; Veil of ignorance (Rawls); Well-being

JEL Classifications
D6; J63

Economics, as a discipline connected to policy decisions, has always been involved in the analysis of social objectives and of their underlying ethical values.

More recently, an important corpus of theories has been developed in which the formal analytical tools of economic theory are mobilized in order to relate basic principles of social ethics to precise criteria for the evaluation of social states of affairs. This corpus is not yet fully unified, and is still replete with debates and open questions, but the outlook of the field is rapidly evolving.

Economics is also connected to individual ethics, as economic decisions by households and firms sometimes involve issues of morality and responsibility toward partners and co-traders. Experimental economics has revealed that altruism, fairness and reciprocity considerations are a key component of strategic interactions between individuals. Ethical issues for firms relate to 'business ethics' and 'corporate social responsibility'.

Normative Versus Positive Economics

'Positive economics' seeks to understand and explain economic mechanisms, whereas 'normative economics' deals with the assessment of policies or states of affairs. While this distinction is useful, one must not be misled into believing that a clear-cut separation is obtained in practice. Positive economics is naturally inclined to focus on ethically important social and economic issues. Moreover, forecasting the consequences of various policies belongs to its realm and, even though this is in principle a purely positive task, it is practically relevant when it is associated with normative conclusions. Meanwhile, normative economics contains many results which merely consist in a logical analysis of the content of given concepts or the relations between concepts. In such work the economist can largely put his own value judgements aside. As Samuelson (1947, p. 220) aptly wrote, 'it is a legitimate exercise of economic analysis to examine the consequences of various value judgments, whether or not they are shared by the theorist'. The role of ethical judgements in economics has received valuable scrutiny in Sen (1987), Hausman and McPherson (2006) and Mongin (2006).

Wariness about value judgements has often led economists to shun issues of distribution and to focus on efficiency as encapsulated in the Pareto principle. In particular it is tempting to think that a Pareto improvement (that is, a new situation that everyone in the population prefers to the status quo) cannot be questioned since, by definition, everyone would approve it. But focusing on Pareto improvements does not protect from controversy. The most pressing reason is that there are important ethical values, especially regarding the distribution, which are not captured by the Pareto principle. As a consequence, there may exist a non-Pareto improving reform that is much better for social welfare than any Pareto improvement. In particular, in the presence of large inequalities, focusing on Pareto improvements implicitly amounts to condoning the status quo, as noted in Arrow (1951). A second reason is that people's immediate preferences, which are relied upon in applications of the Pareto principle, may not correctly represent people's interests. The difficult issue of defining individual interests is examined below.

The Pareto criterion, on the other hand, is very restrictive because it applies only when unanimity of preference is obtained. New Welfare Economics (surveyed in Chipman and Moore 1978) and cost–benefit analysis have developed less restrictive criteria dealing with potential Pareto improvements (that is, reforms that would be Pareto improving if certain transfers from the gainers to the losers were performed in addition, even though they are not actually implemented). Such criteria have been condemned by many authors, from Arrow (1951) to Blackorby and Donaldson (1990), for being inconsistent (they can produce cyclic social preferences), and also

unethical because they may approve reforms that seriously hurt the worst-off sub-populations (because the transfers that could compensate the losers may not be actually made). More sophisticated versions of cost–benefit analysis (Drèze and Stern 1987) rely on consistent and distribution-sensitive social welfare functions.

The Branches of Normative Economics

Four main branches of normative economics can be distinguished. The first – the theory of social choice, sparked off by Arrow (1951) with a provocative impossibility theorem – examines the properties of functions that define an ordering of a set of alternatives (policies, social states, candidates) on the basis of the ordinal non-comparable (ONC) preferences of the population over these alternatives. In economics individual ONC preferences, which can be retrieved from observable choices, are usually considered the natural informational basis. However, the theory of social choice has been extended (following Sen 1970a) to cover the possibility of incorporating more information about individual utilities. This theory has achieved a thorough understanding of the properties of voting rules (surveyed in Brams and Fishburn 2002; Pattanaik 2002) and of the informational requirements about individual utilities of various social welfare functions (surveyed in d'Aspremont and Gevers 2002; Bossert and Weymark 2004). The second – the theory of fair allocation, initiated by Kolm (1972) – studies the allocation of resources among individuals with heterogeneous tastes and abilities, in terms of fairness criteria such as no-envy (no agent should prefer another's bundle), lower bounds (for example, no agent should prefer the equal-split allocation), solidarity (for example, no agent should be hurt by an increase in available resources), among others (see survey in Moulin and Thomson 1997). The third – the theory of inequality and poverty measurement, originally anchored to the Gini coefficient and the Lorenz curve – focuses on income distributions and has developed after Kolm (1969), Atkinson (1970) and Sen (1973) into an axiomatic theory of indices, dominance criteria (that is, criteria ascertaining that a distribution dominates another for a family of indices), and, more recently, statistical tests to be performed on samples. Surveys of this field can be found in Atkinson and Bourguignon (2000) and Silber (1999). The fourth – the theory of axiomatic bargaining and cooperative games, initiated by Nash (1950) and Shapley (1953) – analyses how to find a fair compromise in utility possibility sets, under different assumptions about coalition formation. A synthesis on axiomatic bargaining is available in Thomson (1999).

Three other branches must be mentioned, which can be viewed perhaps as sub-branches of the main ones. Connected to social choice theory, Harsanyi's impartial observer argument (1953) and aggregation theorem (1955), offered in defence of utilitarianism, have generated an important literature and some debates (see Broome 1991; Weymark 1991; Roemer 2002). Sen's (1970b) and Gibbard's (1974) liberal paradoxes have also triggered debates about how to formalize rights and incorporate them in the theory of social choice (see in particular Gaertner et al. 1992; Arrow et al. 1997, ii, part IV). The theory of axiomatic cost and surplus sharing, which lies somewhere between cooperative games and fair allocation, studies the allocation of cost or surplus shares across individuals not as a function of their preferences but as a function of their actions (demands or contributions). It is surveyed in Moulin (2002).

The various branches have developed more or less independently but, if one puts aside the theory of cooperative games and the theory of cost-surplus sharing, they can all be formally described as seeking to rank alternatives of various sets X on the basis of the population's utility functions $U_1, \ldots, U_n$, and possibly other personal characteristics such as abilities and needs. The theory of social choice usually considers only one given set X and in Arrow's initial version, retains information only about individual ONC preferences; the theory of fair allocation takes the Xs to be sets of feasible allocations and usually seeks only to identify a subset of optimal allocations in each X on the basis of ONC preferences; the theory of

inequality and poverty indices usually takes X to be the set of income distributions and focuses on a special aspect of distributions instead of a general notion of social welfare, although it sometimes establishes a link between special indices and social welfare functions; the theory of axiomatic bargaining usually ignores the structure of the sets X and directly examines the utility possibility sets $\{(U_1(x),..,U_n(x))|x \in X\}$. This formal similarity between the various branches is favourable to cross-fertilizing. One observes that the frontiers between these fields, which are largely due to the contingent circumstances of their creation, are progressively vanishing, to be replaced by more substantial differences in principles, such as whether ONC preferences provide morally sufficient information about individual well-being.

Cross-fertilizing with political philosophy is also an essential part of the history of the field. Rawls's (1971) theory of justice, borrowing many features from economics, has had a profound influence in return in at least three ways. It has rekindled interest for equality and the maximin principle among economists; it has popularized the idea that putting individuals behind a 'veil of ignorance' concerning their own circumstances, as in Harsanyi's impartial observer argument (with differences in assumptions and conclusions which have aroused controversies between these two authors and commentators), is a way to define justice; it has also provided an implicit justification of the theory of fair allocation by defining justice in terms of equality of resources (even if by resources he meant 'primary goods', that is, all-purpose goods rather than ordinary commodities), firmly rejecting interpersonal comparisons of utility across individuals with incommensurable preferences. Dworkin's (2000) theory of equality of resources makes a clear reference to the no-envy criterion and combines it with the veil of ignorance in a hypothetical insurance market in which individuals can insure against bad personal characteristics. Social policy should then, according to this theory, mimic the hypothetical insurance premiums and indemnities by suitable taxes and transfers. Sen's (1992) theory of capabilities proposes to shift the focus from resources to functionings, a general notion of individual achievement, and to seek equality of capability sets, that is, the sets of functionings that are accessible to individuals. Arneson (1989) and Cohen (1989) have also proposed to focus on opportunities rather than achievements on the ground that individuals should be viewed as responsible for seizing the opportunities that are offered to them. These recent theories of justice have generated an increased interest in normative economics for the notions of freedom and responsibility (see, for example, Roemer 1998, and, for surveys, Fleurbaey and Maniquet 2006; Peragine 1999; Barberà et al. 2004a, b). Among many other philosophical contributions that have been influential in normative economics, one must also mention Parfit's (1984) thought-provoking essay on utilitarianism, identity and population issues.

The Measurement of Individual Well-Being

With reference to the traditional social welfare function $W(U_1, \ldots, U_n)$, it is convenient to decompose the problem of defining a criterion for the evaluation of social states into two sub-problems, namely, the problem of assessing each individual situation and the problem of constructing a synthetic measure for the whole population. This decomposition, however, does not imply that the former is any less normative than the latter. The measurement of individual well-being is not just an empirical exercise and raises many ethical issues.

First, in such measurement one must consider whether one should take account of individuals' political and social preferences or only of their tastes about their personal situation. It appears that these two kinds of preferences belong to different levels of social evaluation. Political and social preferences are relevant in the democratic debate about general principles, while personal tastes belong to the concrete evaluation of social situations. Mixing the two levels may yield absurd consequences. In particular, making the allocation of resources directly depend on the satisfaction of individual political preferences may produce grossly unfair distributions, for instance when a

simple summation of utilities representing heterogeneous political preferences induces the altruist to transfer their resources to the malevolent. As Sen (1977) nicely put it, one must not confuse the aggregation of 'judgments' with the aggregation of 'interests'. It is worth emphasizing that, concerning judgement aggregation, the contribution of normative economics is not limited to studying voting procedures with the goal of neutrally aggregating judgements, because it may play an active role in shaping those judgements. Indeed, by scrutinizing issues in *interest* aggregation it clarifies the substance of the debates and may help in the formation of personal *judgements* on these matters. It may thereby make a useful contribution in the deliberation process and the construction of a consensus.

When dealing with the aggregation of 'interests', personal tastes themselves are not necessarily appropriate as a basis for ranking social states, since they may be influenced by unjust social pressures and conditioning, or based on mistaken beliefs. It may then appear preferable to try to guess what individual tastes would be if formed in correct conditions. Practical procedures eliciting such authentic preferences have yet to be invented, but one may observe that 'safety belt' policies are commonplace and are usually justified by reference to people's well-considered interests.

A more radical questioning of the reference to personal tastes comes with the observation that, even in absence of conditioning or bad information, those with demanding tastes do not necessarily deserve more resources than those with more modest wishes. Should individuals not assume responsibility, in some cases, for their 'expensive tastes' (Dworkin 2000)? The emergence of principles of freedom and responsibility in recent theories of justice has in fact revealed how important such considerations have always been in the selection of relevant dimensions of individual well-being. Even the standard principle of consumer sovereignty according to which every individual is the best judge of his own interests implies that the social criterion will let the allocation of personal resources be managed under the individual's sole responsibility and will at most cater to a synthetic measure of his satisfaction.

More importantly, if individuals are considered responsible for how they transform ordinal satisfaction into numerical utility, then social evaluation can disregard their utility functions and take care of their ONC preferences only. The focus of Arrovian social choice and of the theory of fair allocation on ONC preferences instead of utilities can hereby find a justification in terms of individual responsibility, in addition to more traditional arguments about the difficulty to compare subjective utility across individuals. Similarly, in Sen's theory of capabilities, the social criterion deals with capability sets and disregards what combinations of functionings individuals choose in those sets. The normative appreciation of what individuals should be held responsible for (or be left free to handle by themselves) and what they should not is a difficult domain of philosophical debating to which economists are not necessarily well equipped to contribute. But economic models are very convenient to examine the consequences of various choices in this matter and it is instructive to relate various policy choices to underlying attributions of responsibility to the target population (see, for example, Roemer 1998).

The important divide between welfarist and non-welfarist approaches is largely connected to this issue. A welfarist approach retains subjective utility (interpreted in terms of happiness or in terms of satisfaction) as the ultimate metric of well-being. A non-welfarist approach will typically discount subjective utility and take account of objective achievements, resources, opportunities or rights, although Sen's theory, for instance, does retain utility as one relevant functioning among others. Critiques of welfarism invoke not only the (at least partial) responsibility of individuals for their subjective utility, but also the idea that there are some objective dimensions of achievement which matter independently of their effect on satisfaction. For instance, a physical disability may justify help even if the concerned individual has perfectly adapted to his situation in terms of subjective utility. Or granting basic rights and freedoms may be viewed as so essential to the constitution of a community of morally

autonomous agents that they should be granted uniformly, independently of their potentially unequal effect on individuals' various utility functions. More generally, fairness in the allocation of resources typically involves non-welfarist concerns. For instance, the axiomatic theory of bargaining, as recalled above, disregards the economic allocations and focuses exclusively on utility possibility sets. This has counter-intuitive consequences. Consider problem 1 which consists in deciding with which probability Ann or Bob will win a ten-dollar prize, and problem 2 which also consists in deciding winning probabilities, except that if Ann wins she gets ten dollars whereas if Bob wins he only gets one dollar. Nash's bargaining solution (which maximizes the product of utility gains) selects the fifty–fifty solution for both problems, whereas it would seem more reasonable to give Bob a greater probability of winning in the second case. See Roemer (1996) for a detailed criticism of welfarism in axiomatic bargaining.

The welfarist–non-welfarist distinction, however, is mainly philosophical and the economist can always reinterpret the utility index U_i as an index of capability, opportunity or objective advantage instead of subjective utility, without changing much to the formal analysis of normative criteria. What is more important for economic analysis is whether the relevant data about individual situations consist in a numerical index permitting comparisons across individuals or in ONC preferences only (a third possibility, considered in the theory of multidimensional inequality, is when individual situations are described by vectors of numerical indices, with no synthetic index or ordering). In the first case, one has a kind of 'formal welfarism' and the standard framework of social welfare functions is readily available. In the second case (as well as the third), one may eventually be able to construct a comparable index, but such an index is not given a priori and its construction must be justified.

The indexing problem is considered a vexing issue for non-welfarist theories of justice such as Rawls's (involving an index of primary goods) or Sen's (involving an index of capabilities). Indeed, it appears that if this index is personalized so as to espouse each individual's preferences, it is then a utility representation of preferences and one is back into the welfarist framework. The alternative is to impose a uniform index to all individuals, but this is tantamount to adopting a special view of how to weight the various goods (or capabilities), that is, a dogmatic or perfectionist definition of the good life. At this point economic analysis is helpful because it shows that a non-welfarist approach can nonetheless respect individual preferences. Consider the simple case of an exchange market with identical prices at the various allocations under consideration. Then, on the assumption that individuals are free to choose their consumption in their budget set, indexing their well-being by the market value of their consumption is congruent with their preferences over consumption bundles, although it is certainly non-welfarist since across individuals there is no relation between utility and wealth. More generally, with each preference ordering one can always associate an index function which is ordinally equivalent to the welfarist measure of utility without coinciding with it, as noted in Roemer (1996). In conclusion, between the pure welfarist approach and the 'perfectionist' non-welfarist approach, there is room for 'Paretian' non-welfarist approaches which respect individual preferences. This distinction can be formally described as follows. In the problem of ranking alternatives on the basis of the profile $(U_1, \ldots, U_n)$, the welfarist approach relies on the utility values $(U_1(x), \ldots, U_n(x))$ at each alternative x; the perfectionist approach ignores individual preferences, imposes an index U^* and evaluates x in terms of $(U^*(x), \ldots, U^*(x))$; the 'Paretian' non-welfarist approach retains the ONC preferences represented by $U_1, \ldots, U_n$ and constructs an ordering of alternatives obeying the Pareto principle (additional examples are provided below).

Another issue for economic analysis is whether the social state must be described in terms of consequences or in terms of procedures. Most of normative economics is still largely consequentialist, but the growing focus on opportunities, rights and freedom of choice definitely enlarges the scope of analysis beyond narrow consequentialism. So far, the studies of rights and freedom

have remained rather abstract, dealing with the general definition of rights, the foundations of a measure of individual freedom and the analysis of distributions of opportunity sets, but there is some interest for more concrete economic settings (on these various approaches, see, for example, Laslier et al. 1998; Pattanaik et al. 2004). Contractarian theories of justice, which analyse justice norms as being shaped by individuals' interests in mutual cooperation, also appeal to game theorists. For instance, Binmore (1994–8) relates various degrees of egalitarianism and libertarianism to different time horizons of social interaction, arguing that the latter prevails in the long run.

The Definition of Social Criteria

When a suitable index of well-being $U_1, .. ., U_n$ is given, as in the welfarist or the perfectionist approaches, the only problem that remains is to choose a social welfare function W from the menu offered by the theory of social choice. For instance, the sum-utilitarian function

$$W(U_1(x), \ldots, U_n(x)) = U_1(x) + \ldots + U_n(x)$$

displays no aversion to inequality in utilities, the maximin function

$$W(U_1(x), \ldots, U_n(x)) = \min\{U_1(x), \ldots, U_n(x)\}$$

has an infinite aversion, while the product (or Nash) function

$$W(U_1(x), \ldots, U_n(x)) = U_1(x) \ldots U_n(x)$$

is an example of an intermediate function. With a small or zero inequality aversion over utilities, as with the utilitarian function, priority is given to individuals with a high marginal utility, independently of their utility level, whereas with a high-inequality aversion, as with the maximin, priority is given to the worst-off (in terms of utility level), even if they have a low marginal utility.

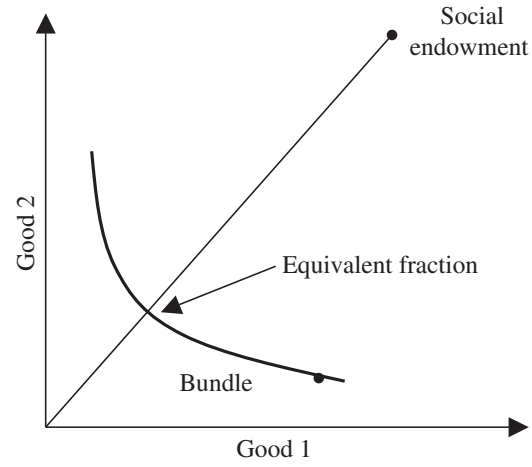
In the choice of an appropriate degree of inequality aversion, it is often thought that the veil of ignorance provides a helpful guide. In the simple version of this device (Harsanyi, Rawls), the observer simply imagines that she could become any individual of the considered society. One may introduce variations about whether this implies taking on all the personal characteristics of each possible individual or simply some of them (his ability, his preferences). In more complex versions of the scheme (Dworkin's hypothetical insurance, or Rawls's original position under some interpretations), individuals may all be put behind the veil and be left free to bargain or make transactions. The attraction of the veil of ignorance comes from the obvious fact that it guarantees the impartiality of decisions. But impartiality is a very weak requirement and any symmetric social welfare function such as those listed above is impartial. The important issue in the choice of a social welfare function is not that it must be impartial, but how averse to inequality it should be. In this respect, the veil-of-ignorance device appears actually ill-suited. It links inequality aversion to the degree of risk aversion of the observer. A very risk-averse observer will come close to some maximin criterion, whereas a risk-neutral observer will adopt a utilitarian kind of rule. If she maximizes her expected utility, her decision will at any rate be structurally utilitarian in some metric. How this translates into a choice of social welfare function W depends on the specific form of the ignorance veil (that is, what personal characteristics are inherited by the observer when she becomes a particular individual). In Dworkin's hypothetical insurance, expected-utility maximizers will adopt insurance contracts that will produce utilitarian kinds of allocations. It is hard to find a reason why risk aversion about the possibility to becoming different persons should determine distributive judgements about actually existing individuals. One could as well imagine other devices, such as living all the lives of the population, one after the other, in a reincarnation process. Again, one does not see why intertemporal preferences over sequences of reincarnations should be especially relevant for distributive judgements over coexisting individuals. The attraction of the veil of ignorance may perhaps come from the mistaken belief that

impartiality is all there is to social justice. But the theory of social choice clearly shows that impartiality is just a minimal requirement. The multiplicity of impartial observer devices (lottery, reincarnation.. .) proves that each of them is just one way among many others to achieve impartiality, and presumably none of them reaches a comprehensive view of justice (on these issues, see, for example, Roemer 2002).

If one forgets the veil and genuinely thinks about inequality aversion in terms of fairness between existing individuals, one still faces a moral dilemma, since typically the social welfare functions that do not give full priority to the worst-off have the repugnant feature that a very small gain to many well-off individuals can always, if these individuals are sufficiently numerous, outweigh a large loss to a badly-off person. On the other hand, the maximin function always prefers giving a very small gain to the worst-off, no matter how costly this may be to all the other individuals. The way out of this dilemma has yet to be invented.

Aggregating Preferences

Let us now turn to the case in which ONC preferences are the only relevant data about individual utilities. Arrow's impossibility theorem of social choice suggests that there is a conflict between the Pareto principle and impartiality in this context, in contrast to the context of the previous paragraph in which many social welfare functions were simultaneously increasing and symmetric in utilities. But this alleged conflict occurs only when one requires the social ranking of two alternatives to depend only on how individuals rank them, to the exclusion of any other alternative. This 'independence' requirement is very restrictive and precludes using information about individual preferences such as their marginal rate of substitution, how they compare their current bundle to natural benchmarks, and so on. The theory of fair allocation actually features many fairness criteria (such as no-envy or lower bounds) which violate this requirement by examining extended portions of indifference curves in order to evaluate an allocation.



Ethics and Economics, Fig. 1

Consider the following example. A certain quantity of divisible goods has to be distributed. At any allocation in which all individuals receive a personal bundle, evaluate every individual's bundle by the fraction of the social endowment that is equivalent according to this individual's preferences (see Fig. 1).

This is a standard way of representing individual preferences by an index function, and such indexes can then serve as arguments of a social welfare function W. This exemplifies a Paretian and impartial way to aggregate individual ONC preferences. This example has an interesting life of its own in the literature. It is briefly examined in Arrow (1951, p. 31), and rejected on the ground that it violates the above independence requirement. This observation, however, could as well suggest abandoning the requirement. A variant of the example is mentioned in Kolm (1969). It is invoked by Samuelson (1977) in order to show that it is possible to construct a Bergson–Samuelson social welfare function on the sole basis of ONC preferences. As explained in Samuelson (1947), such a function is a mapping

$$E(x) = W(U_1(x), \ldots, U_n(x))$$

where $U_1, .. ., U_n$ are suitable indices representing individual preferences. (That the construction depends only on ONC preferences is verified by the fact that the same function E can be written

with other, ordinally equivalent, indices $V_1, .., V_n$ provided W is correspondingly adapted – some commentators have identified W instead of E as the fixed Bergson–Samuelson function, thereby concluding that Samuelson must have been wrong.) Eventually, it is used by Pazner and Schmeidler (1978) in the definition of the concept of egalitarian equivalence, which has become quite important in the theory of fair allocation (an allocation is egalitarian-equivalent if it is Pareto-equivalent to a resource egalitarian allocation).

This example shows a simple way to aggregate preferences: first construct an index and then apply a standard social welfare function. The Borda rule, in the voting context, is another example of the same vein. The selection of the index need not be arbitrary and the above example, for instance, refers to fractions of the social endowment and thereby makes sure that individuals who prefer their bundle to the equal-split are always considered better-off than those in the opposite situation. There are less simple aggregation methods. Consider for instance the index

$$U_i^p(x_i) = e_i(U_i(x_i), p) - p\omega_i,$$

where x_i and ω_i denote i's personal bundle and endowment, e_i his expenditure function, p a price vector. At every feasible allocation x (that is, such that $x_1 + ... + x_n \leq \omega_1 + ... + \omega_n$) and for every price vector p, one has

$$U_1^p(x_1) + \ldots + U_n^p(x_n) \leq \sum_i (px_i - p\omega_i) \leq 0,$$

while a feasible allocation x^* is a competitive equilibrium associated to price vector p^* if and only if one has $U_i^{p*}\left(x_i^*\right) = 0$ for all i. Therefore, for any inequality-averse social welfare function W, the function

$$E(x) = \max_{p \in s} W(U_1^p(x_1), \ldots, U_n^p(x_n)),$$

(where S is the simplex of appropriate dimension) exactly selects the competitive equilibria as the best allocations in the set of feasible allocations. This function, which bears some similarity to cost–benefit criteria without sharing their drawbacks, is slightly more complex than the previous examples as it makes the evaluation of individual situations depend on a price vector that itself depends on the whole allocation. Observe how even the maximin criterion, which is just a particular case of inequality-averse social welfare function, can rationalize a competitive equilibrium, no matter how unequal the endowments are. This is due to the deduction of the value of endowments in the index U_i^p, which is justified if individuals are held responsible for their endowments, so that one is not interested in individual total consumptions but only in the value difference between their consumption and their endowment.

Even though such constructions are based on ONC preferences, they always involve some kind of interpersonal comparison (of the relevant indexes) in order to determine who should be given priority in the allocation of resources (for a synthesis on interpersonal comparisons in general, see Fleurbaey and Hammond 2004).

Hard Issues

A few hard ethical issues have already been described. There are many others. Consider, for instance, Fig. 2. It describes two individuals' utility under three different policies, depending on a random state. Policy B is better than A, since for the same *ex post* distribution of utilities it provides a less unequal distribution of prospects *ex ante*. And Policy C is better than B, since for the same distribution of *ex ante* prospects it guarantees an equal distribution of utilities *ex post*.

However, a social criterion that computes the expected value of social welfare will be indifferent between A and B, while a criterion satisfying the Pareto criterion with respect to expected utilities will be indifferent between B and C. Harsanyi's utilitarian criterion, which satisfies both properties, is indifferent between the three policies. The search for a better criterion in this context is still going on (see, for example, Ben Porath et al. 1997).

Ethics and Economics, Fig. 2

Policy A

	State (or effort) 1	State (or effort) 2
Individual 1	1	1
Individual 2	0	0

Policy B

	State (or effort) 1	State (or effort) 2
Individual 1	1	0
Individual 2	0	1

Policy C

	State (or effort) 1	State (or effort) 2
Individual 1	0	1
Individual 2	0	1

A similar difficulty plagues the theory of equal opportunities, since, in spite of essential differences, there is an obvious similarity between random prospects and opportunities. In the same figure, replace the random states by effort levels exerted by the individuals. Then one can read the rows as depicting opportunity sets for the individuals. Under this new interpretation, Policy B is still better than Policy A because opportunity sets are less unequal, and Policy C is even better because it perfectly equalizes the opportunity sets. None of the social criteria in the literature displays this pattern of preference, because each of them focuses either on the distribution of *ex ante* opportunities or on the *ex post* neutralization of the effect of variables for which the agents are not responsible (this issue is discussed in Fleurbaey and Maniquet 2006).

Another issue is the comparison of social welfare across different populations. The theory of social choice is curiously restricted, in its standard formulation, to the ranking of options for a given population (with a specific ranking for each possible profile of preference of this population). But economic analysis is recurrently asked to compare standards of living across time (measurement of growth) or space (international comparisons). The framework of social choice should then be extended in order to rank not just allocations but pairs (allocation, population) involving populations with different preferences and different sizes. Interestingly, there are contexts in which size should be a neutral matter (for example, comparison of living standards between big and small countries) and other contexts (demographic policy) for which a theory of the optimal size is needed. Optimal demography is a famously hard domain. Classical utilitarianism, which is based on the total sum of utilities, has the unappealing feature that a population with arbitrarily low average welfare is always better, if it is sufficiently large, than any given population (Parfit 1984). The criteria proposed by Blackorby et al. (2005) (see also Broome 2004) avoid this 'repugnant conclusion' by computing individual welfare as the surplus $U_i - U^*$ over some positive threshold of utility U^* and then apply a social welfare function to these surpluses. The U^* threshold corresponds to the minimal welfare level that an individual must reach in order for his addition to society to be an improvement. Such criteria may thus induce judgements that a given population would be better off without its members whose utility is below the threshold, even when these

members have positive utility. There is again a dilemma here.

This is just a sample of those hard ethical issues about social evaluation that economic analysis may never be able to render easy, but is able to clarify and to which it does, sometimes, give inventive solutions.

See Also

- ▶ Arrow, Kenneth Joseph (Born 1921)
- ▶ Bargaining
- ▶ Bergson, Abram (1914–2003)
- ▶ Cost–Benefit Analysis
- ▶ Egalitarianism
- ▶ Equality of Opportunity
- ▶ Fair Allocation
- ▶ Harsanyi, John C. (1920–2000)
- ▶ Inequality (Measurement)
- ▶ Libertarianism
- ▶ Pareto Principle and Competing Principles
- ▶ Philosophy and Economics
- ▶ Samuelson, Paul Anthony (1915–2009)
- ▶ Sen, Amartya (Born 1933)
- ▶ Social Choice (New Developments)
- ▶ Social Contract
- ▶ Social Welfare Function
- ▶ Voting Paradoxes

Bibliography

Arneson, R. 1989. Equality and equal opportunity for welfare. *Philosophical Studies* 56: 77–93.

Arrow, K. 1951. *Social choice and individual values*. 2nd ed. New Haven: Yale University Press 1963.

Arrow, K., A. Sen, and K. Suzumura, eds. 1997. *Social choice re–examined*, 2 Vols. International Economic Association New York: St Martin's Press. London: Macmillan.

Arrow, K., A. Sen, and K. Suzumura, ed. 2002–2006. *Handbook of social choice and welfare*. Amsterdam: North-Holland.

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Atkinson, A., and F. Bourguignon, ed. 2000. *Handbook of income distribution*. Vol. 1. Amsterdam: North-Holland.

Barberà, S., W. Bossert, and P. Pattanaik. 2004a. Ranking sets of objects. In *Handbook of utility theory*, ed. S. Barberà, P. Hammond, and C. Seidl, Vol. 2. Dordrecht: Kluwer.

Barberà, S., P. Hammond, and C. Seidl, ed. 2004b. *Handbook of utility theory*. Vol. 2. Dordrecht: Kluwer.

Ben Porath, E., I. Gilboa, and D. Schmeidler. 1997. On the measurement of inequality under uncertainty. *Journal of Economic Theory* 75: 194–204.

Binmore, K. 1994–8. *Game theory and the social contract. Vol. 1: Playing fair. Vol. 2: Just playing*. Cambridge: MIT Press.

Blackorby, C., and D. Donaldson. 1990. Review article: The case against the use of the sum of compensating variations in cost–benefit analysis. *Canadian Journal of Economics* 23: 471–494.

Blackorby, C., W. Bossert, and D. Donaldson. 2005. *Population issues in social-choice theory, welfare economics and ethics*. Cambridge: Cambridge University Press.

Bossert, W., and J. Weymark. 2004. Utility in social choice. In *Handbook of utility theory*, ed. S. Barberà, P. Hammond, and C. Seidl, Vol. 2. Dordrecht: Kluwer.

Brams, S., and P. Fishburn. 2002. Voting procedures. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.

Broome, J. 1991. *Weighing goods: Equality, uncertainty, and time*. Oxford: Basil Blackwell.

Broome, J. 2004. *Weighing lives*. Oxford: Oxford University Press.

Chipman, J., and J. Moore. 1978. The new welfare economics, 1939–1974. *International Economic Review* 19: 547–584.

Cohen, G. 1989. On the currency of egalitarian justice. *Ethics* 99: 906–944.

d'Aspremont, C., and L. Gevers. 2002. Social welfare functionals and interpersonal comparability. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.

Drèze, J., and N. Stern. 1987. The theory of cost–benefit analysis. In *Handbook of public economics*, ed. A. Auerbach and S. Feldstein, Vol. 2. Amsterdam: North-Holland.

Dworkin, R. 2000. *Sovereign virtue: The theory and practice of equality*. Cambridge, MA: Harvard University Press.

Fleurbaey, M., and P. Hammond. 2004. Interpersonally comparable utility. In *Handbook of utility theory*, ed. S. Barberà, P. Hammond, and C. Seidl, Vol. 2. Dordrecht: Kluwer.

Fleurbaey, M., and F. Maniquet. 2006. Compensation and responsibility. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.

Gaertner, W., P. Pattanaik, and K. Suzumura. 1992. Individual rights revisited. *Economica* 59: 161–177.

Gibbard, A. 1974. A Pareto-consistent libertarian claim. *Journal of Economic Theory* 7: 388–410.

Harsanyi, J. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434–435.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Hausman, D., and M. McPherson. 2006. *Economic analysis, moral philosophy, and public policy (2nd edn of Economic analysis and moral philosophy)*. Cambridge: Cambridge University Press.
Kolm, S. 1969. The optimal production of social justice. In *Public economics*, ed. J. Margolis and H. Guitton. London: Macmillan.
Kolm, S. 1972. *Justice et équité*, Paris: Ed. du CNR-S. Translated as *justice and equity.* Cambridge, MA: MIT Press, 1997..
Laslier, J., M. Fleurbaey, N. Gravel, and A. Trannoy, ed. 1998. *Freedom in economics: New perspectives in normative analysis*. London: Routledge.
Mongin, P. 2006. Value judgments and value neutrality in economics. *Economica* 73: 257–286.
Moulin, H. 2002. Axiomatic cost and surplus sharing. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.
Moulin, H., and Thomson, W. 1997. Axiomatic analysis of resource allocation problems. In *Social choice re–examined*, 2 vols. International Economic Association New York: St Martin's Press, ed. K. Arrow, A. Sen, and K. Suzumura. London: Macmillan, vol. 1.
Nash, J. 1950. The bargaining problem. *Econometrica* 18: 155–162.
Parfit, D. 1984. *Reasons and persons*. Oxford: Clarendon Press.
Pattanaik, P. 2002. Positional rules of collective decision-making. In *Handbook of social choice and welfare*, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.
Pattanaik, P., M. Salles, and M. Suzumura, eds. 2004. Non welfaristic issues in normative economics. Special issue of *Social Choice and Welfare* 22(1).
Pazner, E., and D. Schmeidler. 1978. Egalitarian equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92: 671–687.
Peragine, V. 1999. The distribution and redistribution of opportunity. *Journal of Economic Surveys* 13: 37–69.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Roemer, J. 1996. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.
Roemer, J. 1998. *Equality of opportunity*. Cambridge, MA: Harvard University Press.
Roemer, J. 2002. Egalitarianism against the veil of ignorance. *Journal of Philosophy* 99: 167–184.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P. 1977. Reaffirming the existence of 'reasonable' Bergson–Samuelson social welfare functions. *Economica* 44: 81–88.
Sen, A. 1970a. *Collective choice and social welfare*. San Francisco: Holden Day.
Sen, A. 1970b. The impossibility of a Paretian liberal. *Journal of Political Economy* 78: 152–157.
Sen, A. 1973. *On economic inequality*. Oxford: Clarendon Press (expanded edn, Sen, A. and Foster, J., 1997).
Sen, A. 1977. Social choice theory: A re-examination. *Econometrica* 45: 53–90.
Sen, A. 1987. *On ethics and economics*. Oxford: Basil Blackwell.
Sen, A. 1992. *Inequality re-examined*. Oxford: Clarendon Press.
Shapley, L. 1953. A value for N-person games. In *Contributions to the theory of games*, ed. H. Kuhn and A. Tucker, Vol. 2. Princeton: Princeton University Press.
Silber, J., ed. 1999. *Handbook of income inequality measurement*. Dordrecht: Kluwer.
Thomson, W. 1999. *Bargaining theory: The axiomatic approach*. New York: Academic Press.
Weymark, J. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal comparisons of well-being*, ed. J. Elster and J. Roemer. Cambridge: Cambridge University Press.

EU European Semester

Marco Scipioni
Post-Doctoral Fellow, Johns Hopkins University School of Advanced International Studies Bologna Center, Bologna, Italy

Abstract

The European Semester is an EU effort to better coordinate member states' economic policies. Coordination of national fiscal and economic policy has long been considered essential for the tenability of a monetary union lacking the federal features of taxing and spending. The European Semester coordinates national economies by setting collective deadlines for the submission, analysis and discussion of national budgetary plans as well as the various forms of economic coordination that the EU has accumulated through time, from the Macroeconomic Imbalance Procedure (an agreement amongst EU members to prevent risky macroeconomic policy) to Europe 2020 (the EU's growth strategy).

In policy terms, the European Semester provides more evidence for those who saw the Economic and Monetary Union (EMU) developing through soft coordination, as some of the most notable outputs are Council

recommendations addressed to member states. However, more intrusive tools such as the Stability and Growth Pact (the agreement by which all EU member states commit to maintaining the value of the euro through fiscal responsibility) are now framed within the Semester schedule. Importantly, the design of this framework exemplifies a broader shift in EU economic governance from an ex post monitoring of member states' preferences to an effort to ex ante shape their adoption. This occurs by scheduling EU discussions before national ones, thus aiming to inform and shape the latter, and by imposing synchronised submission of different policy documents, thus forcing member states to be consistent in their commitments across policy areas. This might raise questions as to the limits of EU action, particularly in the light of subsidiarity concerns.

While the Semester is now part of the official jargon of EU and member states' economic policy, commitment to such processes in the absence of external pressures (e.g. crisis) has been questioned, as recorded by falling implementation rates. Further, while the first years of the Semester were associated with weak economic growth, the picture has recently turned more positive. That said, there seems to be limited effect on macroeconomic imbalances, one of the main stated objectives of the framework.

Keywords

EMU governance; Monitoring and coordination; EU integration

JEL Codes

E61; F02; O52

Related Articles

Begg: Genuine Economic and Monetary Union. http://www.dictionaryofeconomics.com/article?id=pde2014_E000339&edition=current&q=genuine%20economic%20and%20monetary%20union&topicid=&result_number=1.

Warin: Stability and Growth Pact. http://www.dictionaryofeconomics.com/article?id=pde2008_S000488&edition=current&q=stability%20and%20growth&topicid=&result_number=1.

Larch and Jonung: Stability and Growth Pact of the European Union, The http://www.dictionaryofeconomics.com/article?id=pde2014_S000556&edition=current&q=stability%20and%20growth&topicid=&result_number=8.

Wyplosz: European Monetary Union http://www.dictionaryofeconomics.com/article?id=pde2008_E000300&edition=current&q=european%20monetary%20union&topicid=&result_number=2.

The European Semester represents the most recent EU effort to better coordinate member states' economic policies. Coordination of national fiscal and economic policy has long been considered essential for the tenability of a Union lacking the federal features of taxing and spending. This is crystallised in the Treaties, where reference is made to the need for member states to 'regard their economic policies as a matter of common concern' and to 'coordinate them within the Council' (European Union 2012: Art. 121), and for euro area countries to 'strengthen the coordination and surveillance of their budgetary discipline' and 'to set out economic policy guidelines for them, while ensuring that they are compatible with those adopted for the whole of the Union and are kept under surveillance' (European Union 2012: Art. 136). Academic commentators have described policy coordination as 'supranational rules or norms which are agreed by all member states, leave primary responsibility for the policy area with national authorities, but set limits on their discretion' (Begg et al. 2003: 66). The rationale and extent of such coordination, beyond the narrow fiscal policy targets included in the Maastricht Convergence Criteria (that is, the criteria which EU member states need to meet in order to be accepted to the euro), has been subject to much debate. Alesina et al. argued 'explicit coordination of monetary and fiscal policy is not necessary, if the monetary and fiscal authorities (independently) follow appropriate

and prudent policies' (Alesina et al. 2001: 7). Issing reinforced this point by adding that close coordination between monetary and fiscal authorities risks blurring respective competences and therefore weakening accountability mechanisms (Issing 2002). Eichengreen and Wyplosz argued that efforts at consolidating budget implied by the Stability and Growth Pact (SGP) (the agreement by which all EU member states commit to maintaining the value of the euro through fiscal responsibility) would consume politicians' political capital to tackle structural reforms, significantly limiting their growth prospects as well as their capacities to jointly stimulate the economy – in other words, more budgetary surveillance might entail less economic coordination (Eichengreen and Wyplosz 1998). Collignon highlighted that, considering that economic policies are kept at the state level, 'one cannot expect that purely independent decisions will necessarily result in the efficient provision of collective goods', and hence might become detrimental to a common undertaking such as the single market (Collignon 2001). Pisani-Ferry further pointed out that in a decentralised system such as the EMU, 'where decision makers are numerous and diverse [. . .] discretionary co-ordination inevitably implies high transaction costs' (Pisani-Ferry 2002: 9), thus revealing the added benefit of coordination.

In practice, the European Semester aims at coordinating national economies by putting collective deadlines on the submission, analysis and discussion of national budgetary plans as well as the various forms of coordinated monitoring that the EU has accumulated through time. It is a 'semester' as after a period of EU-level coordination (approximately from November to May each year), a period of national coordination follows.

This entry first puts the European Semester in the context of the efforts to build and strengthen economic coordination at the EU level. Second, it looks at continuities and changes in the Semester as compared to previous economic coordination exercises. Third, it provides a description of the European Semester and places it in the context of the EMU reforms undertaken since 2010. Fourth, it looks at the early assessments of its results. The fifth section concludes the entry.

The Rationale Behind the European Semester

Scholars and policy-makers alike have intensely debated the nature of the relationships between monetary policy on the one hand and fiscal and economic policies on the other, in the context of Economic and Monetary Union (EMU). Early policy proposals on monetary union in the 1970s regarded these two poles as inseparable and foresaw nothing short of 'budget aggregates' decided at the EU level (Pisani-Ferry 2006: 824–825). Since the creation of EMU, monetary policy has been fully delegated to the European Central Bank (ECB), member states have cooperated on budgetary policy with binding rules envisaging an institutional hierarchy as well as monitoring and sanctioning powers at the EU level (the SGP), while other economic policies (e.g. structural reforms, employment policies) have been subject to a mixture of policy coordination forms where no sanctions were in order and policy outputs that mainly materialised in recommendations (e.g. the Broad Economic Policy Guidelines, BEPGs: non-legally binding recommendations guiding annual economic policy in member states).

Once a decision to create a monetary union is taken, there remains a choice for either a centralised or decentralised budget that can automatically stabilise the euro-area (De Grauwe 2016). Due to political reasons (Hodson 2009), a centralised budget has not been feasible in the EU. That leaves only the national budgets to stabilise business cycle shocks (De Grauwe 2016: 214–217). Here, policy-makers hit a conundrum. On the one hand, fiscal policies remain their only tool to deal with negative shocks. On the other, unconstrained budgetary policies represent a risk for the monetary union in terms of negative spillovers in interest rates from highly indebted countries to others and regarding ECB independence. Begg et al. (2003: 67–68) summarise the political economy rationale for economic coordination as

aiming to solve two problems. First, negative externalities are likely to stem from excessive government deficits in the context of currency union, which are then spread onto other member states in the form of a 'higher unified interest rate' (for a critique, see Eichengreen and Wyplosz 1998: 76–77). The second relates to the provision of independence and ultimately the credibility of the central bank in the context of a multiplicity of different fiscal policies. If markets presume that the central bank will ultimately come to rescue states with high deficit and unsustainable public debt, its credibility is at risk. To solve both these issues, cooperation and coordination provide some form of collective insurance (Schelkle 2005).

The EMU solution to the dilemmas posed by its institutional configuration was to establish some form of pre-accession convergence (the Maastricht Convergence Criteria), to tie continued membership of the EMU to some form of adherence to budgetary rules (the excessive deficit procedure within the Treaty and its further specification within the SGP), and to coordinate economic policies (see Treaty obligations in Art. 2(3), Art. 5(1) TFEU). Since Maastricht, the EU has consistently decided to reduce economic divergence and the creation of excessive budgetary deficits in some countries for the purpose of common interest. It has done so by setting EU-wide rules, and at the same time issuing guidelines to nudge member states towards commonly agreed objectives for macroeconomic policies and structural reforms.

Economists diverge as to how effective such a solution was. Indeed, besides the 'lack of planning to deal with financial stability' and 'lack of transparency of the ECB', Giavazzi and Wyplosz identify 'the poor articulation between monetary policy and national fiscal policy' as one of the three main flaws in the construction of the EMU (2015: 723–724). De Grauwe (2010c: 3) holds that such asymmetric institutional architecture (centralised for monetary policy, decentralised for economic ones) has been problematic since its inception. Further, such an asymmetrical arrangement 'disregards elementary principles of political economy' inasmuch as it confounds the appropriate level where political power should be exercised (De Grauwe 2010b: 3). While monetary policy has been fully supranationalised, De Grauwe states powers over tax and spending should remain at the national level because that is where accountability mechanisms are grounded. Such arrangements do not and cannot work, De Grauwe argues, simply because states have no incentives to comply and draw their legitimacy from their electorates, not EU institutions. The emphasis on ownership that emerged in EU official documents in the mid-2000s (see next section), can be regarded as a surrogate to overcome such fundamental political economy problems. The basic idea is that the more governments buy into the reform process, the less likely they will be to back down from that path in the face of organised interests or electorate opposition. Confirming this importance, Directorate-General for Economic and Financial Affairs (DG ECFIN) Director Marco Buti has recently declared 'the take-away, from my perspective, is not that fiscal rules per se proved their inadequacy within the crisis. Rather, the main lesson is that fiscal rules operate in a certain institutional and political environment, which has to be conducive to sensible implementation' (Buti 2016: 186).

A Brief History

The first manifestations of economic coordination were the Broad Economic Policy Guidelines (BEPGs), included in the Treaty of Maastricht, and now based on Article 121 TFEU. The BEPGs' main objective was to maintain and foster economic convergence, thus limiting the potential impact of asymmetric economic cycles and shocks in the context of a currency union. In terms of scope, the BEPGs provided recommendations to member states on structural reforms and macroeconomic policies, 'ranging from budgetary policy and wage developments to labour market reform and financial-market integration' (Deroose et al. 2008: 828). The alignment of the Lisbon Agenda (the agreement amongst EU member states to make the EU the leading knowledge-driven economy by 2010) into the

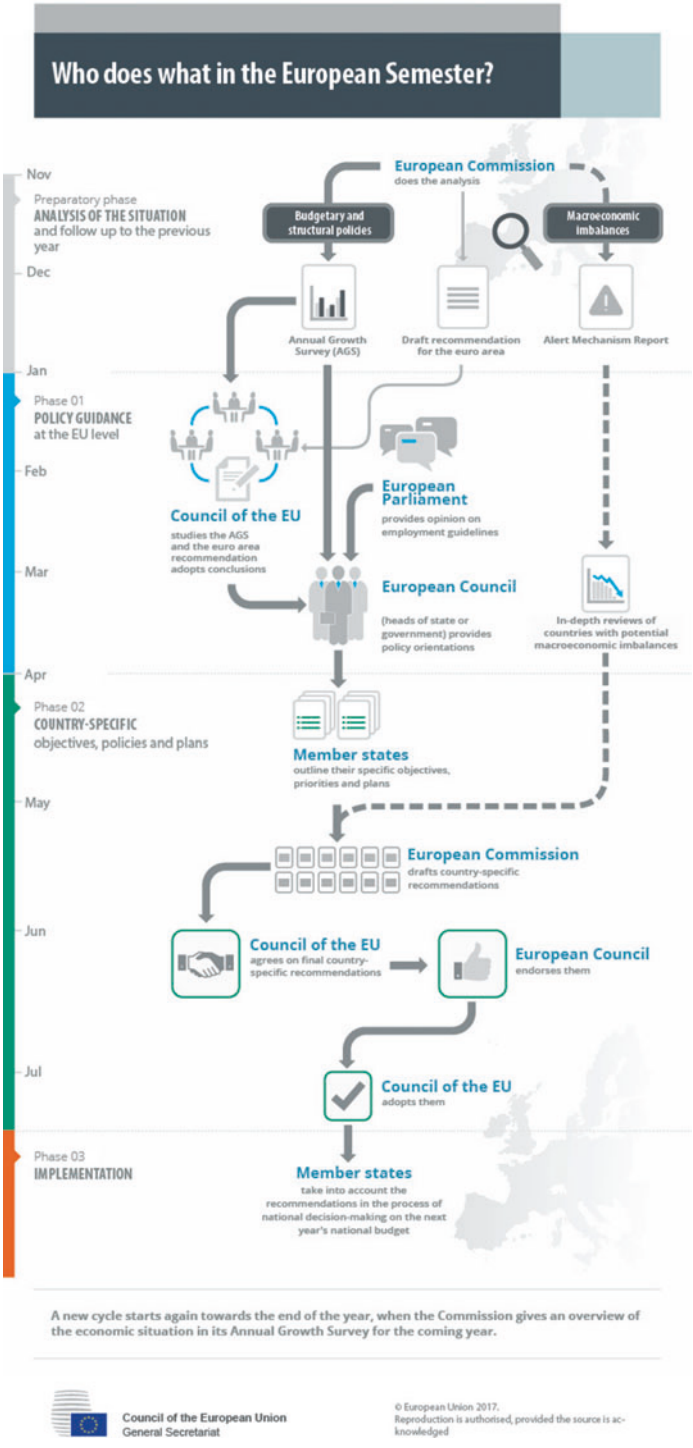
BEPGs in 2000 provoked a significant expansion of the scope of issues considered, as well as the perception that their all-encompassing nature was to the detriment of priority and focus, overcomplicating the process, and ultimately diluting its success (Deroose et al. 2008). The absence of peer pressure amongst member states and a general lack of commitment further contributed to the widespread perception that the BEPGs had limited effect (Deroose et al. 2008). This was paralleled by a negative assessment of the Lisbon Strategy (Kok 2004).

The European Semester itself was first proposed in a 2004 Commission's Communication wherein the opportunity represented by the failure of SGP implementation was seized upon to suggest 'a revision of the economic policy calendar' (Commission 2004: 7). Notably, to strengthen the coordination between member states' economies and to facilitate the monitoring role of EU bodies, the Commission proposed an 'EU semester' to be followed by a national one where budgets would be prepared. Such sequencing would have allowed 'the BEPGs and the Opinions on the [Stability and Convergence] programmes to be taken into account in the preparation of national budgets by governments' (Commission 2004: 7). In the eyes of the Commission, coordination under the Semester would have had several merits: reorienting the SGP-related programmes from 'description' to 'strategic planning'; increasing the national 'ownership' of EU economic coordination as interactions between the two levels intensified; strengthening legitimacy because of the deeper involvement of national parliaments in the policy-making process. Since these early proposals it has been possible to notice a gradual shift towards consensus-building before measures are enacted, and away from what is increasingly perceived as ultimately toothless monitoring. Ownership issues were contemplated at the time, and are still paramount in the Semester (Zuleeg 2015). The focus on ownership derives from IMF past experiences in structural reform, which indicated that the degree to which local policy and political elites buy into reform programmes is determinant to their success (Bird and Willett 2004).

How Does It Work?

The crisis that exploded in Europe in 2010 alerted policy-makers to the need to equip the Union with a single framework to coordinate all the activities related to the reformed EU economic governance. The coordination through the Semester of the correction of macroeconomic imbalances, unsustainable budgetary positions, as well as competitiveness and growth strategies is an answer to the lessons of the crisis as well as the longer-term EU project of coordinating *en bloc* member states' growth strategies. The crisis has revealed how an exclusive focus on a reduced basket of indicators (i.e. mainly budget deficits) was not conducive to macroeconomic stability. Having a framework where a larger pool of indicators is discussed and possibly coordinated was regarded as a step forward from past policies.

Officially started in January 2011, the European Semester cycle kicks off between November and December each year when the Commission releases the Annual Growth Survey (AGS), which spells out the priorities to meet the jobs and growth agenda of Europe 2020, and targets the EU as a whole for the following year (Fig. 1). The AGS builds on three elements: an evaluation of Europe 2020 Strategy, the Macroeconomic Report and the Joint Employment Report. At the same time, the Alert Mechanism Report (AMR) is released, which is part of the constant monitoring of the Macroeconomic Imbalance Procedure (MIP: an agreement amongst EU members to prevent risky macroeconomic policy) (Commission 2014: 7). In addition, the Commission issues opinions on draft budgetary plans. These opinions are then discussed in the Council, while the European Parliament (EP) is involved in the Economic Dialogue on discussions over the AGS. In December/January the Commission and member states have bilateral meetings, while euro-area member states approve their budgets. In January, the Commission can carry out inspections in the member states to verify anomalies that emerged during monitoring. At the same time, the Council adopts recommendations on the euro-area and conclusions on the AGS and AMRs. In March, the European Council



EU European Semester, Fig. 1 The European semester. http://www.consilium.europa.eu/en/policies/european-semester/ (© European Union, 1995–2017)

EU European Semester, Table 1 Number of CSRs by legal basis

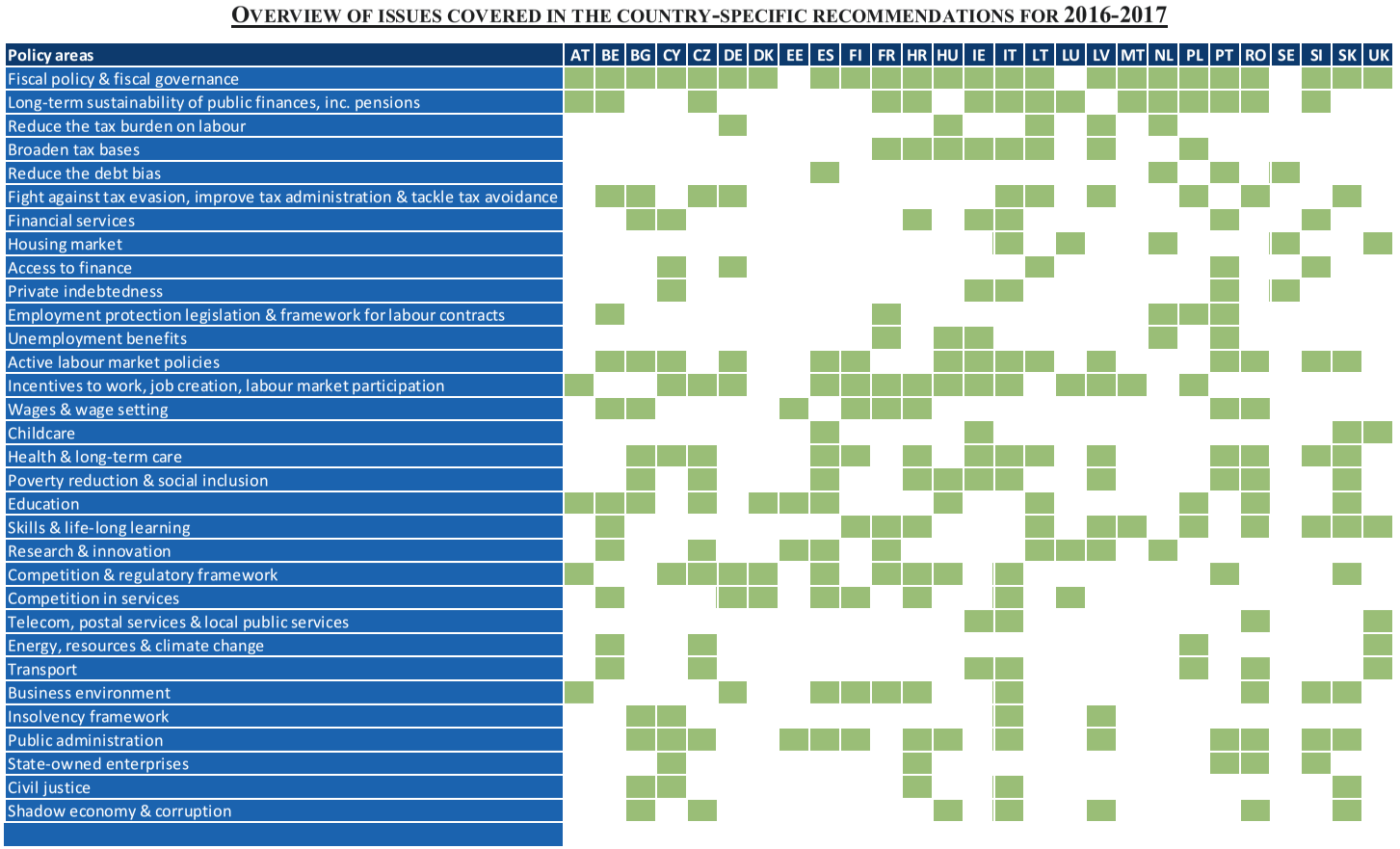
European semester	Exclusively SGP		Exclusively MIP		Jointly SGP and MIP		Integrated guidelines		Total
2012	18	(13%)	31	(22%)	5	(4%)	84	(61%)	138
2013	18	(13%)	50	(35%)	6	(4%)	67	(48%)	141
2014	19	(12%)	58	(37%)	8	(5%)	72	(46%)	157
2015	11	(11%)	48	(47%)	10	(10%)	33	(32%)	102
2016	13	(15%)	36	(40%)	9	(10%)	31	(35%)	89

Source: EP 2016a, © European Union, 2016. http://www.europarl.europa.eu/RegData/etudes/ATAG/2014/528767/IPOL_ATA(2014)528767_EN.pdf

adopts the economic priorities on the basis of the AGS. Following suit, the Commission again has bilateral meetings with member states which focus on their submission of National Reform Programmes (NRPs: the local implementation plan of EU objectives) and Stability and Convergence Programmes (SCPs: the individual country's budgetary plans for the coming years, connected with the member states' obligations under the SGP). The National Reform Programmes are compiled by the member states under their Europe 2020 commitments and should be in line with the BEPGs and EGs as well as the Annual Growth Survey (AGS). In May, the Commission releases the Country Specific Recommendations (CSRs) which encompass both economic and budgetary policies and draw on member states' performance for the previous year. At this point, the Commission starts discussions with the EP. Between June and July the Council first debates the CSRs and then the European Council endorses them. The cycle comes to an end in September, when euro area member states submit to the Commission the draft budgets for the following year, while in October the EP discusses the new AGS. The submission of draft budgets should allow the Commission to check their consistency with commitments made in the Stability Programmes previously submitted. It should be noted that there is a differentiation at the heart of the process between euro- and non-euro members, reflecting the various arrangements present in the instruments that the Semester encapsulates (e.g. SGP, MIP). Finally, countries that are under specific surveillance programmes are excluded from the regular CSRs activities, as is currently the case for Greece, and was for Cyprus (between 2013 and 2015), Ireland (2012–2013), and Portugal (2012–2013).

As already mentioned, the Semester has an encompassing nature, which derives from hosting basically every mechanism that the EU has developed so far regarding fiscal, budgetary and economic coordination (see Table 1 below for a recent overview). For instance, the 2015 European Semester started with an AGS focused on three pillars: 'boost for investment' (connected with the other flagship project of an Investment Plan for Europe, which focuses on investment in infrastructure), 'structural reforms', and 'fiscal responsibility'. Examples of specific CSRs are: stepping up efforts to fight tax fraud and evasion in Spain and Italy, increasing tax revenues by harmonising VAT in Germany and introducing tax incentives to strengthen home ownership in the Netherlands (EP 2014). What is the added value of synchronising such a wide array of policy issues? Ideally, to lead member states to a consistent approach under the several coordination schemes present in the EU which cut across policy areas. At a minimum, this means coherence between structural reforms and their budgetary implications. In parallel, synchronisation would facilitate effective monitoring by magnifying inconsistencies between policies (Fig. 2).

As shown in Table 1 below, the numbers of SGP-related recommendations have slightly decreased in absolute terms but remain fairly stable as a share of the total; the MIP-related too have remained fairly stable in absolute numbers but, due to the decreasing overall numbers, have increased in overall share; the most dramatic decrease, however, is in the Integrated Guidelines (which set out the framework for the Europe 2020



EU European Semester, Fig. 2 Overview of issues covered in the EU CSRs for 2016–2017. http://ec.europa.eu/europe2020/pdf/csr2016/csr2016-overview-table_en.pdf (© European Union, 1995–2017)

strategy and reforms at Member State level), which have more than halved in absolute terms and nearly so in relative terms.

As pointed out by the EP, failure to implement recommendations might result in different kinds of sanction, depending on the legal basis of the individual recommendations. For this reason, it may be misleading to encapsulate the European Semester as either soft or hard law, as the framework displays features of both (EP 2016a). Although finding an explicit legal recognition in EU secondary legislation (the Six Pack), this 'overarching framework' (to use the Commission's words; 2014: 10) is only as legally binding and only envisages the possibility of sanctions (two elements frequently used to distinguish soft/hard law policies) as its components are.

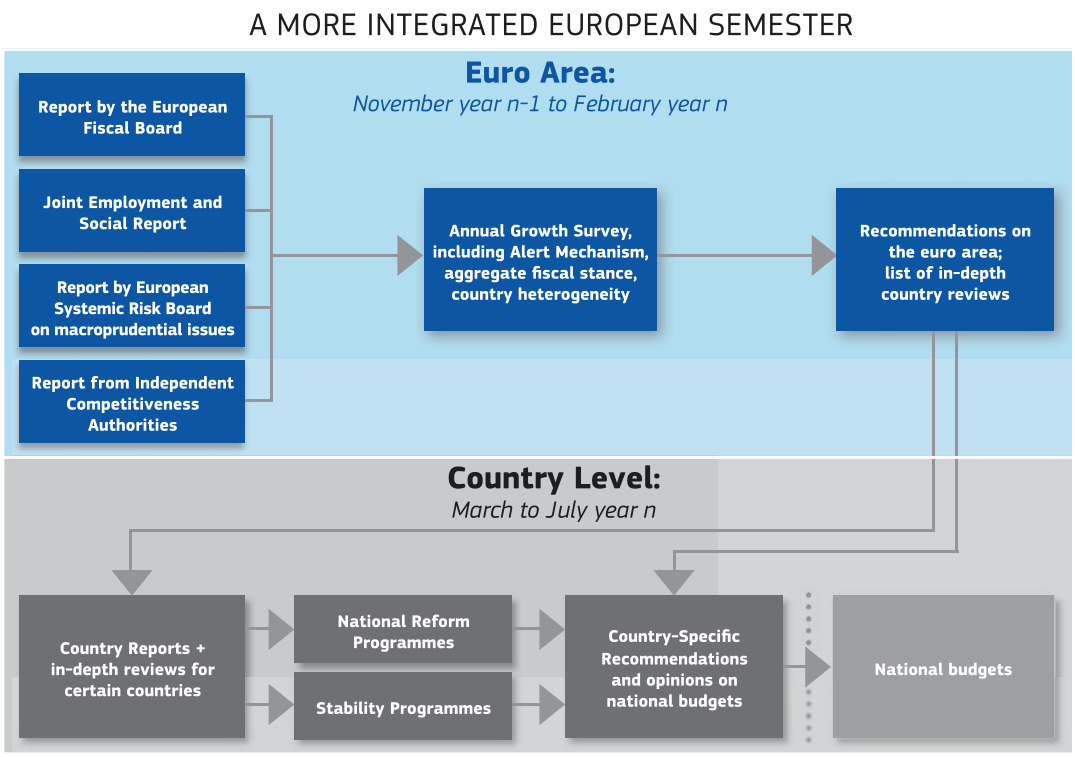
Finally, the Semester can be conceived as having two sources of legitimacy. The first is the involvement of the EP through the Economic Dialogue (albeit in a loose fashion), which might be regarded as enhancing legitimacy by the sheer fact of being directly elected. The second is the fact that, after a European Semester, a national semester is envisaged, where member states are supposed to internally debate the measures to be adopted by their parliaments. The Commission has also recently tightened deadlines during the EU-level of the Semester with the stated aim of conferring more time to national-level discussion on economic reforms and budgets.

The Track Record So Far

Looking at implementation, commentators agree that the picture is disappointing (Darvas and Leandro 2015; Gros and Alcidi 2015; Zuleeg 2015). Member states have made 'full' or 'substantial' progress in decreasing shares of the total recommendations addressed to them between 2012 and 2014 (from 12.3% to 6.4%), while the percentage of measures that were either only limitedly or not implemented increased from 28.8% to 48.1% (Gern et al. 2015: 8). In other words, there is an increasing risk of irrelevance for the entire exercise. Other studies have quantified the level of implementation, finding that 'reform efforts' had been spent only on 40% of recommendations in 2011 (Darvas and Leandro 2015: 10), decreasing to 29% by 2014. Euro-area member states, subject to tighter coordination and monitoring, performed only limitedly better, and for instance in 2014 implemented 31% of recommendations, as compared to 23% of countries outside the Eurozone. Further, implementation is analogous to what was achieved under a parallel OECD framework, which casts doubt on the merits of an operation which has become the flagship initiative for economic coordination in the EU (Darvas and Leandro 2015: 11–13).

Commentators have adduced several reasons for such a deficient implementation, including weak institutional architecture, softening of external pressures as the crisis abated, lack of national ownership of the reform agenda, unfeasible and poorly specified economic objectives. Such implementation gaps have pushed some economists to argue that 'procedures culminating in sanctions in the case of persistent non-compliance' should be introduced to safeguard the 'functioning of the EMU as a whole' (Gern et al. 2015: 24). However, it should be noted that experience with the emphasis on sanctioning as a deterrent for non-compliance has been mixed in the case of SGP (De Grauwe 2010a; Hodson 2014), and it is not clear why this case should be different. Others have regarded such implementation gaps as the ultimate evidence of the need for a 'systemic governance reform' and called for, *inter alia*, a 'fiscal capacity' that member states could tap into provided that they met 'contractual' obligations to be devised by the Commission in the context of the European Semester (Zuleeg 2015: 12, 14). However, the empirical evidence for such a move towards more binding coordination is mixed. Comparing recommendations belonging to the SGP and those of the MIP and other policies (the former two are legally binding, not the latter), studies have found that while the older and more tightly coordinated SGP marginally outperform the others, the 44% implementation rate on average for the period 2012–2014 is not cause for celebration (Darvas and Leandro 2015: 13–14).

In 2015, the Semester was reshuffled into a 'Revamped European Semester' after taking



EU European Semester, Fig. 3 A more integrated European Semester. https://ec.europa.eu/priorities/sites/beta-political/files/5-presidents-report_en.pdf (© European Union, 1995–2017)

stock of mixed implementation records (Commission 2015). EU officials seem to have converged on the idea that less is better, meaning that addressing fewer recommendations to member states might be easier for EU bodies to monitor and for national administrations to implement (Gern et al. 2015: 5). The reforms aim at creating even tighter integration for euro-area member states, as well as gaining more political traction at the domestic level (with an eye to the unresolved issue of ownership). Institutionally, two new advisory bodies are created – a new European Fiscal Board (EFB) and a system of Competitiveness Authorities (Fig. 3).

Commentators have highlighted that while frequent reference is made to 'euro-area aggregate fiscal stance', it is not clear how an 'optimal aggregate fiscal stance should be determined' (Darvas and Leandro 2015: 6). The emphasis on the aggregate fiscal stance comes from ECB invitations and is conceived by the Commission's officials as 'the sum of country policies' (Buti 2016). The Commission argues that there is a 'clear sub-optimal repartition of the fiscal adjustment across countries', but 'those who do not have fiscal space want to use it; those who have fiscal space do not want to use it' (EP 2016b: 2). However, there seems to be little official indication as to how this adjustment should take place (EP 2016b), and policy instruments are still lacking for 'stabilising economic cycles' or forcing 'countries to run larger budget deficits' (Darvas and Leandro 2015). The recommendation to Germany to use the available fiscal space to boost investment was the only divisive one between member states sitting in the Council and the Commission in 2016 (Council 2016), but a decisive one in terms of helping alleviate the internal imbalances of the Union. This reflected (and re-confirmed) the difficulty of EU coordination to push for a measure that was widely regarded as collectively beneficial in the face of

domestic opposition. Further, Deroose and Griesse observe a higher rate of implementation in 'areas where market pressure requires an imminent policy response' (Deroose and Griesse 2014: 1, 6), which casts doubt on whether the positive effects of the Semester are due to coordination or external pressure.

Looking at 2017–2018, the Commission has recently forecast a neutral stance until 2018. In its opinion, member states should run an expansionary fiscal policy, while also taking into account the ECB monetary policy (EP 2016b). Giavazzi regards the very concept of euro area aggregate fiscal stance as a response to the ECB hitting one of the limits of monetary policy, meaning the decision to impose a zero percent interest rate (Giavazzi 2016). Tabellini argues that the crisis has shown that the current setting of the EU, where full financial integration as well as stability is sought for but there is no common fiscal policy or tools to manage aggregate demand, is untenable (Tabellini 2015). In this vein, the emphasis on the aggregate fiscal stance would probably be regarded as ineffective, as one of the main lessons from the crisis was that the EU needs nothing short of a fiscal union to provide 'fiscal stabilisation for the euro-area as a whole' and 'resources to withstand systemic financial crisis' (Tabellini 2016). While a euro-area aggregate fiscal stance that targets, for instance, 'country-specific demand deficiencies' might be adequate to deal with fiscal stabilisation, its suitability to deal with systemic financial crisis is more contested (Monacelli 2016; Wyplosz 2011). In any case, this neglects the crucial question of how such coordination could come about, taking into consideration that coordination initiatives so far have yielded mixed results.

While the Commission seems positive regarding the implementation of the European Semester (Darvas and Leandro 2015: 25), others have outlined a series of criticisms. A line of criticism that often emerges in commentaries relates to the unclear selection criteria upon which recommendations are made, or prioritised (EP 2012: 9). The basic point of a cycle was that recommendations could be followed up year on year. However, commentators have noticed an unevenness of the themes featuring in successive waves of recommendations. For instance, social and labour market policies have been maintained in some countries but not others from one cycle to another despite ostensible lack of progress in all of them (Gern et al. 2015). Others have observed an overall reduction in recommendations for social inclusion, education and skills, without an apparent justification (Zuleeg 2015: 11). Others have criticised the lack of focus on specific areas with significant potential for negative spillovers, such as financial sector vulnerabilities and environmental issues (Gern et al. 2015: 12), or standards on tax compositions (Darvas and Leandro 2015: 21).

Whether the Semester contributes to a more sustainable EMU is complex to assess, but looking at some fundamental macroeconomic variables since the start of the framework might provide an initial indication. Euro area GDP growth increased to 2% in 2015 after weak if not negative performance since 2011, particularly when compared to international competitors such as the USA. Euro area unemployment rates decreased in 2015 to 10.9 after an upward trend in the preceding 3 years. Euro area member states continued to consolidate their public balances (reaching −2.1% in 2015) and brought down their public debt after years of an upward trajectory. These aggregate indicators, however, mask significant disparities and varying trends within the Eurozone, which jeopardise EMU sustainability (Hodson 2016: 154–157). It is in this highly uneven context that commentators have wondered why, in contrast with all previous years, 'a recommendation on the need for symmetric intra-euro adjustment [. . .] was not included in the 2015 recommendations' (Darvas and Leandro 2015). DG ECFIN forecasts limited progress until 2018, with four countries still with unemployment rates above 10% (down from six), GDP growth constant and deficits still above the SGP criteria in two countries (down from three) (Commission 2016b: 1; for a similar analysis, see OECD 2016a). Recent studies have shown that core-periphery dynamics of unsynchronised business cycles which marked the run-up to the adoption of the euro have softened but remain present for a subset of member states (Spain, Portugal, Ireland and Greece)

(Campos and Macchiarelli 2016). Considering that remedying these asymmetries was one of the main goals of economic coordination, as it is regarded as one of the main factors contributing to making the EMU unsustainable, this suggests that coordination has had limited effects so far.

The first years of operation have confirmed that national governments are reluctant to fully buy into the process of coordination (Gern et al. 2015: 5). This shows that lack of ownership, to which the Semester should have been an answer, is still at the forefront. More transparency and timely information on implementation have been suggested as possible solutions with a view to increasing the take up at national level. According to commentators (Gern et al. 2015: 6), increased transparency could be achieved by making more explicit the theoretical premises upon which recommendations to member state are made. More involvement of national parliaments is regarded as instrumental in delivering ownership, but implementation so far suggests that efforts at better including them have largely failed (Gern et al. 2015: 9). Other commentators have pointed out that the only way out from the current situation of lack of ownership and the consequent implementation gap, as well as sovereign and legitimacy issues, is nothing less than an 'economic government', even though there is ample recognition that this 'is not on the cards politically' (Zuleeg 2015: 7). The EP has tried to increase ownership by raising awareness of the process, for instance by inviting national finance ministers to hearings in the EP's Economic and Monetary Affairs Committee as part of Economic Dialogues set up by the Semester (EP 2014).

Conclusions

As the overarching framework including both hard and soft EMU policies, the European Semester is noteworthy for its design. First, by scheduling EU policy discussions before national ones it aims to inform and shape the latter, and within the Eurozone this takes an intrusive form because of, *inter alia*, the budgetary implications of the Two Pack (which involves closer monitoring of the deficits of euro area member states). Second, by imposing synchronised submission of different policy documents, it forces member states to be consistent in their commitments across policy areas, which in turn facilitates monitoring by EU institutions. This exemplifies a broader shift in EU economic governance from an ex post monitoring of member states' preferences to an effort to ex ante shape their adoption. In this regard, it should be remembered that the Semester, by including a wide range of mechanisms such as Europe 2020 and the MIP, now covers a broad spectrum of issues, ranging from energy policy to civil justice. Besides the Semester, the Commission 'considers that ex ante coordination should concern only major national economic reform plans and that it should take place at an early stage before the measures are adopted' (Commission 2013: 3). Because of such methods, timing and scope, the Semester raises questions as to the limits of EU action, introducing a degree of tension between economic policy coordination and subsidiarity concerns.

To conclude, the Semester should be judged by its outcomes. Two simple questions might help in this assessment. Has the Semester facilitated the achievement of the SGP, MIP and Europe 2020 objectives? Has it made monitoring more effective? The answer seems to be negative to the former, as implementation of recommendations related to all three regimes has been falling since the start of the process – this in the context of decreasing numbers of recommendations, which in turn helps answer the second question. This fact suggests that policy-makers are converging on the idea that having less demanding monitoring duties is better for both national and European administrations. It also reduces the complexity of the process and avoids duplications. Eventually, it is hoped that focusing on a smaller number of priorities will improve ownership of the reforms, even if experiences in this regard have been disappointing so far. Regarding coordination, the take-home message is that, while the principle of timely and close cooperation has now entered the official jargon of EU and member states' economic policy, commitment to such processes in the absence of external pressures (e.g. the euro area crisis) has been questioned (Gros and Alcidi 2015). This

resonates with evidence from outside the EU of low responsiveness to coordination recommendations in the absence of external pressures, as revealed by the limited success of the OECD advice to some member states (e.g. Austria, Finland, France, Germany, the Netherlands) to undertake further expansionary fiscal policies to support growth than were currently planned (OECD 2016b, c).

According to the Commission, the innovations in EMU governance since 2010 – including the Semester – aim at helping 'foster growth convergence and the achievement of the goals of the Europe 2020 strategy [,] and preventing the build-up of large macroeconomic imbalances'. Regarding growth convergence, the Commission has recently pointed out that 'GDP growth in the euro area has remained slow compared to past recoveries and is not expected to pick up in the coming years', and that 'the cumulated growth of GDP since the end of the recession differs substantially across Member States' (Commission 2016b: ix, 4). As a consequence, the overarching objective of the Union, as stated in the Treaty, of 'balanced economic growth' does not appear particularly strengthened by coordination through the Semester. Regarding Europe 2020, one should note that, despite the reduction in the numbers of recommendations, this has not led to more compliance. Finally, regarding macroeconomic imbalances (EP 2016c), the Commission this year looked sanguine by noticing that 'fewer Member States have economic imbalances than a year ago' (Commission 2016a). However, since 2012 the countries featuring excessive imbalances have increased year after year (Italy and Hungary have been in that category since 2014), while the Commission has been confirming imbalances in Finland, Germany, Ireland, the Netherlands, Spain and Sweden since at least 2014.

Blanchard et al. observed that the more complex agreements are, the harder respective gains are to identify, and the more measures to be taken by all actors differ and have varying contribution to the collective goal, the more difficult effective coordination is to achieve (Blanchard et al. 2013). The Semester might be suffering from similar problems, as with the multiplication of policy instruments, the extended remit of policies under consideration, and the complexity reached by some of the components of the Semester (SGP, MIP), it is possible to argue that it is becoming increasingly difficult for member states to identify the individual gains from such exercises, understand others' contributions and have a clear sense of the overall benefit of the exercise.

References

Alesina, Alberto, et al. 2001. *Defining a macroeconomic framework for the Euro area*, Monitoring European Central Bank. Vol. 3. Geneva: CEPR.

Begg, Lain, Hodson, Dermot, Maher, Imelda. 2003. Economic policy coordination in the European Union. *National Institute Economic Review,* (183).

Bird, Graham, and Thomas D. Willett. 2004. IMF conditionality, implementation and the new political economy of ownership. *Comparative Economic Studies* 46 (3): 423–450.

Blanchard, Olivier, Ostry, Jonathan, Ghosh, Atish. 2013. Obstacles to international macro policy coordination. *VOXEU*. http://voxeu.org/article/obstacles-international-macro-policy-coordination. Accessed 20 Dec.

Buti, Marco. 2016. What future for rules-based fiscal policy? In *Progress and confusion. The state of macroeconomic policy*, ed. Olivier Blanchard et al. London: The MIT Press.

Campos, Nauro, Macchiarelli, Corrado. 2016. A new measure of economic asymmetries in the Eurozone | VOX, CEPR's Policy Portal. *VOXEU*. http://voxeu.org/article/new-measure-economic-asymmetries-eurozone. Accessed 19 Oct.

Collignon, Stefan. 2001. *Economic policy coordination in EMU: Institutional and political requirements*. Harvard: Center for European Studies (CES).

Commission. 2004. Communication from the Commission to the Council and the European Parliament. Strengthening economic governance and clarifying the implementation of the Stability and Growth Pact. COM (2004) 581 final. Brussels.

Commission. 2013. Towards a deep and genuine economic and monetary union ex ante coordination of plans for major economic policy reforms. COM (2013) 166. Brussels.

Commission. 2014. Communication from the Commission to the European Parliament, the Council, the European Central Bank, the European Economic and Social Committee and the Committee of the Regions. Economic governance review. Report on the application of Regulations (EU) n° 1173/2011, 1174/2011, 1175/2011, 1176/2011, 1177/2011, 472/2013 and 473/2013. COM(2014) 905. Brussels.

Commission. 2015. Annual growth survey. Strengthening the recovery and fostering convergence. COM(2015) 690. Brussels.

Commission. 2016a. European semester 2016: Fewer member states have economic imbalances than a year ago. IP-16-591. Strasbourg.

Commission. 2016b. European economic forecast autumn 2016. 38. Brussels.

Council. 2016. Explanations of modifications to Commission Recommendations for the Country Specific Recommendations. ST 9327/16, (updated 13 June 2016). http://data.consilium.europa.eu/doc/document/ST-9327-2016-INIT/en/pdf. Accessed

Darvas, Zsolt, and Álvaro Leandro. 2015. *Economic policy recommendations in the euro area under the European Semester – IPOL_STU(2015)542680_EN.pdf*. Brussels: EP.

De Grauwe, Paul. 2010a. Why a tougher stability and growth pact is a bad idea. *VOX*.

De Grauwe, Paul. 2010b. *What kind of governance for the eurozone?* Vol. 214. Brussels: CEPS.

De Grauwe, Paul. 2010c. *Crisis in the eurozone and how to deal with it*. Vol. 204. Brussels: CEPS.

De Grauwe, Paul. 2016. *Economics of monetary union*. 11th ed. Oxford: Oxford University Press.

Deroose, Servaas, and Jörn Griesse. 2014. *Implementing economic reforms – are EU Member States responding to European Semester recommendations?* Vol. 37. Brussels: Commission, DG ECFIN.

Deroose, Servaas, Dermot Hodson, and Joost Kuhlmann. 2008. The broad economic policy guidelines: Before and after the re-launch of the Lisbon strategy*. *JCMS: Journal of Common Market Studies* 46 (4): 827–848.

Eichengreen, Barry, and Charles Wyplosz. 1998. The stability pact: More than a minor nuisance? *Economic Policy* 13 (26): 65–113.

EP. 2012. An assessment of the European semester, eds. Mark Hallerberg, Benedicta Marzinotto, and Guntram B. WOLFF. Brussels.

EP, 2014. European semester: Why it matters. Updated 20 Jan 2014. http://www.europarl.europa.eu/pdfs/news/expert/background/20121019BKG54051/20121019BKG54051_en.pdf. Accessed 13 Oct 2016.

EP. 2016a. The legal nature of country specific recommendations – IPOL_ATA(2014)528767. Brussels.

EP. 2016b. The Euro area fiscal stance – IPOL_BRI(2016)587374_EN.pdf. Brussels.

EP. 2016c. Implementation of the macroeconomic imbalance procedure – state of play. IPOL_IDA(2016)497739_EN.pdf. Brussels.

European Union. 2012. Consolidated versions of the treaty on European Union and the treaty on the functioning of the European Union. 55. Brussels.

Gern, Klaus-Jürgen, Nils Jannsen, and Stefan Kooths. 2015. *Economic policy coordination in the euro area under the European Semester – IPOL_IDA(2015)542678*. Brussels: EP.

Giavazzi, Francesco. 2016. *Euro-area fiscal stance: Definition, implementation and democratic legitimacy – IPOL_IDA(2016)574426_EN.pdf*. Brussels: EP.

Giavazzi, Francesco, and Charles Wyplosz. 2015. EMU: Old flaws revisited. *Journal of European Integration* 37 (7): 723–737.

Gros, Daniel, and Cinzia Alcidi. 2015. *Economic policy coordination in the euro area under the European Semester – IPOL_IDA(2015)542679_EN.pdf*. Brussels: EP.

Hodson, Dermot. 2009. EMU and political union: What, if anything, have we learned from the euro's first decade? *Journal of European Public Policy* 16 (4): 508–526.

Hodson, Dermot. 2014. Policy-making under economic and monetary union: Crisis, change, and continuity. In *Policy-making in the EU*, ed. Helene Wallace. Oxford: Oxford University Press.

Hodson, Dermot. 2016. Eurozone governance: From the Greek drama of 2015 to the five presidents' report. *JCMS: Journal of Common Market Studies* 54: 150–166.

Issing, Otmar. 2002. On macroeconomic policy co-ordination in EMU. *Journal of Common Market Studies* 40 (2): 345–358.

Kok, Wim. 2004. *Facing the challenge. The Lisbon strategy for growth and employment. Report from the high level group*. Brussels: Commission.

Monacelli, Tommaso. 2016. Asymmetries and Eurozone policymaking. *VOX, CEPR's Policy Portal*. http://voxeu.org/article/asymmetries-and-eurozone-policy making. Accessed 12 Feb.

OECD. 2016a. Developments in individual OECD and selected non-member economies. http://www.oecd.org/eco/outlook/economic-forecast-summary-euro-area-oecd-economic-outlook-november -2016.pdf. Accessed.

OECD. 2016b. OECD economic outlook, vol. 2016(1). Paris.

OECD. 2016c. OECD economic outlook, vol. 2016(2). Paris.

Pisani-Ferry, Jean. 2002. Fiscal discipline and policy coordination in the Eurozone: Assessment and proposals.

Pisani-Ferry, Jean. 2006. Only one bed for two dreams: A critical retrospective on the debate over the economic governance of the euro area*. *JCMS: Journal of Common Market Studies* 44 (4): 823–844.

Schelkle, Waltraud. 2005. The political economy of fiscal policy co-ordination in EMU: From disciplinarian device to insurance arrangement*. *JCMS: Journal of Common Market Studies* 43 (2): 371–391.

Tabellini, Guido. 2015. The main lessons to be drawn from the European financial crisis | VOX, CEPR's Policy Portal. *VOX, CEPR's Policy Portal*. http://voxeu.org/article/main-lessons-be-drawn-european-financial-crisis. Accessed 7 Sept.

Tabellini, Guido. 2016. Which fiscal Union? Building common fiscal policy in the Eurozone. *VOX, CEPR's Policy Portal*. http://voxeu.org/article/building-common-fiscal-policy-eurozone. Accessed 24 Apr.

Wyplosz, Charles. 2011. Eurozone leaders still don't get it. *VOX, CEPR's Policy Portal*. http://voxeu.org/article/eurozone-leaders-still-don-t-get-it. Accessed 25 Oct.

Zuleeg, Fabian. 2015. *Economic policy coordination in the euro area under the European semester – IPOL_IDA(2015)542677*. Brussels: EP.

Eucken, Walter (1891–1950)

Josef Molsberger

Keywords

Economic order vs. economic process; Eucken, W.; Freiburg school; German historical school; Laissez-faire; Social market economy

JEL Classifications

B31

Head of the Freiburg School of German neoliberalism and founder of the yearbook *Ordo,* Eucken was born at Jena on 17 January 1891. Eucken earned his doctoral degree at Bonn (1913). After the *Habilitation* in Berlin (1921), he was professor of economics at Tübingen (1925) and Freiburg (1927–50). He died on 20 March 1950 in London, during a lecture series at the London School of Economics.

Eucken's works mark the return to (neo)classical theory in German economics after the dominance of the Historical School. He stressed, however, the theorist's task to explain reality and rejected model-building if it was purely an intellectual game. Eucken's outstanding analytical contributions include a masterly explanation of the German inflation and currency depreciation on quantity-theoretical grounds (1923), a capital theory (1934) building on Böhm-Bawerk and Wicksell and, in particular, his theory of economic systems (1940) and of economic policy (1952).

Eucken's theory of economic policy starts from the distinction between the *economic order*, the legal and institutional framework of economic activity, and the *economic process,* the daily transactions of economic agents. Under laissez-faire the state neither shapes the economic order nor intervenes in the economic process; in a centrally planned economy the state dominates both. Eucken conceived a *Wettbewerbsordnung* (competitive system) different from both systems: Government should abstain from directly intervening into market processes, but it has to shape the economic order by guaranteeing, through *Ordnungspolitik,* the 'constituent principles' of the market economy (monetary stabilization, free entry, private property, freedom of contract, liability, consistency in economic policy and, primarily, maintaining competition). Subsidiary are the 'regulatory principles': monopoly regulation, social policy, process stabilization policy. Eucken's theory laid the ground for West Germany's 'social market economy'.

Selected Works

1923. Kritische Betrachtungen zum deutschen Geldproblem. Jena: Gustav Fischer.

1932. Staatliche Strukturwandlungen und die Krise des Kapitalismus. *Weltwirtschaftliches Archiv* 36: 297–323.

1934. *Kapitaltheoretische Untersuchungen.* Jena: Gustav Fischer. 2nd enlarged ed. Tübingen: Mohr; Zürich, Polygraphischer Verlag. 1954.

1940. *Die Grundlagen der Nationalökonomie.* Jena: Gustav Fischer. 8th ed. Berlin, Heidelberg and New York: Springer. 1965. Trans. as *The Foundations of Economics.* London. 1950.

1948a. On the theory of the centrally administered economy: An analysis of the German experiment. *Economica* 15; Pt I*,* May, 79–100; Pt II, August, 173–93.

1948b. Das ordnungspolitische Problem. *Ordo* 1: 56–90.

1949. Die Wettbewerbsordnung und ihre Verwirklichung. *Ordo* 2: 1–99.

1951. *Unser Zeitalter der Misserfolge*: *5 Vorträge zur Wirtschaftspolitik. Tübingen: Mohr.*

1952. *Grundsätze der Wirtschaftspolitik,* ed. Eucken, E., Hensel, K.P. Bern: Francke. Tübingen: Mohr. 5th ed. Tübingen: Mohr. 1975.

Bibliography

Böhm, F. 1950. Die Idee des Ordo im Denken Walter Euckens. *Ordo* 3: xv–1xiv.

Jöhr, W.A. 1950. Walter Euckens Lebenswerk. *Kyklos* 4 (4): 257–278.

Lenel, H.O. 1975. Walter Euckens ordnungspolitische Konzeption, die wirtschaftspolitische Lehre in der Bundesrepublik und die Wettbewerbstheorie von Heute. *Ordo* 26: 22–78.

Lutz, F.A. 1961. Eucken, Walter. In *Handwörterbuch der Sozialwissenschaften,* vol. 3. Stuttgart/Tübingen/Göttingen: Fischer/Mohr/Vandenhoeck & Ruprecht.

Schmidtchen, D. 1984. German 'Ordnungspolitik' as institutional choice. *Zeitschrift für die Gesamte Staatswissenschaft* 140 (1): 54–70.

von Stackelberg, H. 1940. Die Grundlagen der Nationalökonomie. *Weltwirtschaftliches Archiv* 51 (2): 245–286.

Welter, E. 1965. Walter Eucken. In *Lebensbilder grosser Nationalökonomen: Einführung in die Geschichte der Politischen Öknonomie*, ed. H.C. Recktenwald. Cologne/Berlin: Kiepenheuer & Witsch.

Euler Equations

Jonathan A. Parker

Abstract

An Euler equation is a difference or differential equation that is an intertemporal first-order condition for a dynamic choice problem. It describes the evolution of economic variables along an optimal path. It is a necessary but not sufficient condition for a candidate optimal path, and so is useful for partially characterizing the theoretical implications of a range of models for dynamic behaviour. In models with uncertainty, expectational Euler equations are conditions on moments, and thus directly provide a basis for testing models and estimating model parameters using observed dynamic behaviour.

Keywords

Calculus of variations; Continuous-time models; Differential equations; Discrete-time models; Dynamic programming; Euler equations; Expectations; Generalized method of moments; Lagrange multipliers; Liquidity constraints; Optimal control; Precautionary saving; Ramsey model; Shadow pricing; Uncertainty

JEL Classifications

C61; E1

An Euler equation is an intertemporal version of a first-order condition characterizing an optimal choice as equating (expected) marginal costs and marginal benefits.

Many economic problems are dynamic optimization problems in which choices are linked over time, as for example a firm choosing investment over time subject to a convex cost of adjusting its capital stock, or a government deciding tax rates over time subject to an intertemporal budget constraint. Whatever solution approach one employs – the calculus of variations, optimal control theory or dynamic programming – part of the solution is typically an Euler equation stating that the optimal plan has the property that any marginal, temporary and feasible change in behaviour has marginal benefits equal to marginal costs in the present and future. On the assumption that the original problem satisfies certain regularity conditions, the Euler equation is a necessary but not sufficient condition for an optimum. This differential or difference equation is a law of motion for the economic variables of the model, and as such is useful for (partially) characterizing the theoretical implications of the model for optimal dynamic behaviour. Further, in a model with uncertainty, the expectational Euler equation directly provides moment conditions that can be used both to test these theoretical implications using observed dynamic behaviour and to estimate the parameters of the model by choosing them so that these implications quantitatively match observed behaviour as closely as possible.

The term 'Euler equation' first appears in text-searchable JSTOR in Tintner (1937), but the equation to which the term refers is used earlier in economics, as for example (not by name) in the famous Ramsey (1928). The mathematics was developed by Bernoulli, Euler, Lagrange and others centuries ago jointly with the study of classical dynamics of physical objects; Euler wrote in the 1700s 'nothing at all takes place in the universe in which some rule of the maximum . . . does not appear' (Weitzman 2003, p. 18). The

application of this mathematics in dynamic economics, with its central focus on optimization and equilibrium, is almost as universal. As in physics, Euler equations in economics are derived from optimization and describe dynamics, but in economics variables of interest are controlled by forward-looking agents, so that future contingencies typically have a central role in the equations and thus in the dynamics of these variables.

For general, formal derivations of Euler equations, see calculus of variations or dynamic programming. This article illustrates by means of example the derivation of a discrete-time Euler equation and its interpretation. The article proceeds to discuss issues of existence, necessity, sufficiency, dynamics systems, binding constraints and continuous-time. Finally, the article discusses uncertainty and the natural estimation framework provided by the expectational Euler equation.

The Euler Equation

Consider an infinitely-lived agent choosing a control variable (c) in each period (t) to maximize an intertemporal objective: $\sum_{t=1}^{\infty} \beta^{t-1} u(c_t)$ where $u(c_t)$ represents the flow payoff in t, $u' > 0$, $u'' > 0$, and β is the discount factor, $0 < \beta < 1$. The agent faces a present-value budget constraint:

$$\sum_{t=1}^{\infty} R^{1-t} c_t \leq W_1 \tag{1}$$

where R is the gross interest rate ($R = 1 + r$ where r is the interest rate) and W_1 is given.

By the theory of the optimum, if a time-path of the control is optimal, a marginal increase in the control at any t, dc_t, must have benefits equal to the cost of the decrease in $t + 1$ of the same present value amount, $-Rdc_t$:

$$\beta^{t-1} u'(c_t) dc_t - \beta^t u'(c_{t+1}) R dc_{t+1} = 0.$$

Reorganization gives the Euler equations

$$u'(c_t) = \beta R u'(c_{t+1}) \text{ for } t = 1, 2, 3 \ldots \tag{2}$$

This set of Euler equations consists of nonlinear difference equations that characterize the evolution of the control along any optimal path. We considered a one-period deviation; several period deviations can be considered, but they follow from sequences of one-period deviations and so doing so does not provide additional information ($u'(c_t) = \beta^2 R^2 u'(c_{t+2})$). These equations imply that the optimizing agent equalizes the present-value marginal flow benefit from the control across periods.

The canonical application of this problem is to a household or representative agent: call c consumption, u utility, and let $W_1 = \sum_{t=1}^{\infty} R^{1-t} y_t$, the present value of (exogenous) income, y. In this case, Eq. (2) imply the theoretical result that variations in income do not cause consumption to rise or fall over time. Instead, marginal utility grows or declines over time as $\beta R 1$; for $\beta R = 1$, consumption is constant.

Existence, Necessity and Sufficiency

In general, to ensure that the Euler equation characterizes the optimal path, one typically requires that the objective is finite (in this example, $u' > 0$) and that some feasible path exists.

Further, since Euler equations are first-order conditions, they are necessary but not sufficient conditions for an optimal dynamic path. Thus, theoretical results based only on Euler equations are applicable to a range of models. On the other hand, the equations provide an incomplete characterization of equilibria. In the example, only by using the budget constraint also can one solve for the time path of consumption; its level is determined by the present value of income.

Dynamic Analysis

More generally, complete characterization of optimal behaviour uses the Euler equation as one equation in a system of equations. For example,

replacing the budget constraint (Eq. (1)) with the capital-accumulation equation

$$K_{t+1} = f(k_t) - c_t + (1-\delta)k_t \qquad (3)$$

where k is capital, $f(k)$ is output, $f' > 0$, $f'' < 0, f(0) = 0$, $\lim_{k \to 0} f_1 > \beta^{-1} - (1-\delta)$, and $\lim_{k \to \infty} f_1 < \beta^{-1} - (1-\delta)$, and adding the constraints k_1 given, $k_t \geq 0$, and $c_t \geq 0$, gives the basic Ramsey growth model. The constant real interest rate of Eq. (2) is replaced by the marginal product of capital in the resulting Euler equation

$$u'(c_t) = \beta(1 - \delta + f'(k_{t+1}))u'(c_{t+1}). \qquad (4)$$

Equations (3) and (4) form a system of two differential equations with two steady states that has been widely studied as a model of economic growth. Linearization shows that the interesting ($k > 0$) steady state is locally saddlepoint stable, and there is a unique feasible convergence path that pins down the dynamic path of consumption and capital.

Binding Constraints

The above Euler equations are interior first-order conditions. When the economic problem includes additional constraints on choice, the resulting Euler equations have Lagrange multipliers. Consider adding a 'liquidity constraint' to our example: that the household maintain positive assets in every period s:

$\sum_{t=1}^{s} R^{1-t} y_t - \sum_{t=1}^{s} R^{1-t} c_t \geq 0$ for all s. In this case, the program is more easily solved in a recursive formulation. Equation (2) holds with a single Lagrange multiplier, $\lambda_{t+1} \geq 0$, on the constraint that assets are positive in $t+1$ since prior to $t+1$ assets levels are unaffected by the choice of c_t and in period $t+1$ the present value of future consumption is unchanged by the one-period deviation considered:

$$u'(c_t) = \beta R u'(c_{t+1}) + \lambda_{t+1}.$$

The multiplier λ_{t+1} has the interpretation of a shadow price. When the constraint does not bind, $\lambda_{t+1} = 0$, the interior version of the Euler equation holds, and the marginal-benefit-marginal-cost interpretation is straightforward. When the constraint binds, the interpretation still holds, but almost tautologically: the change in utility of an extra marginal unit of consumption in t is equal to the change in utility from the marginal decreases in consumption in $t+1$ plus the shadow price (in terms of marginal utility) of marginally relaxing the constraint on c_t. For example, if $\beta R = 1$ and $y_t = \bar{y}\ \ \forall t \neq 2$ and $y_2 = y < \bar{y}$, then $\lambda_{t+1} = 0 \quad \forall t \neq 2\ ,\ \lambda_3 = u'\left(\frac{y+R\bar{y}}{1+R}\right) - u'(\bar{y}) > 0$, and $c_1 = c_2 = \dfrac{y + R\bar{y}}{1+R}, c_t = \bar{y} \quad \forall t \geq 3$. This example illustrates that, relative to the unconstrained equilibrium $(c_t = \bar{y} - r(\bar{y} - y))$, the constraint can postpone consumption ($t = 1, 2$ relative to $t \geq 3$), create a causal link from an increase in income to consumption ($t = 2$ to 3), and can lower consumption in unconstrained periods ($t = 1$).

Continuous Time

In general, continuous-time models have differential Euler equations that are equivalent to the difference-equation versions of their discrete-time counterparts. In the example, replacing $t+1$ with $t + \Delta t$, $c_{t+\Delta t} = c_t + \Delta t$, expanding $u'(c_t + \Delta c_t)$ around c_t, and letting $\Delta t \to 0$ gives:

$$\frac{\dot{c}_t}{c_t} = \sigma_t(r + 1 - \beta)$$

where $\sigma_t = -\frac{u'(c_t)}{c_t u''(c_t)}$. While the marginal-costs-marginal-benefit interpretation of the equation is less obvious in continuous time, it is still clear that consumption rises or falls with the difference between the interest rate (r) and the discount rate ($\beta - 1$), and more obvious that the strength of this response is governed by σ_t, which for this reason is called the elasticity of intertemporal substitution.

Generalized Euler Equations

Dynamic games can also lead to 'generalized' Euler equations. For example, Harris and Laibson (2001) considers a modification of the example as a game among agents at different times who disagree because their preferences are not time consistent due to hyperbolic discounting. At any s, an agent has objective: $u(c_t) + \beta\sum_{\tau=1}^{\infty}\delta^{\tau}u(c_{s+r})$, where $0 < \delta < 1$. Defining recursively $W_{t+1} = R(W_t - c_t)$, the generalized Euler equation is

$$u'(c_t) = R''Effectivediscount\ factor''\left[\underbrace{\beta\delta\left(\frac{\partial c_{t+1}(W_{t+1})}{\partial W_{t+1}}\right) + \delta\left(1 - \frac{\partial c_{t+1}(W_{t+1})}{\partial W_{t+1}}\right)}\right]u'(c_{t+1}).$$

where $c_{t+1}(W_{t+1})$ is the optimal consumption choice made in $t + 1$ as a function of W_{t+1}. The effective discount rate is a function of the (endogenous) marginal propensity to consume wealth in $t + 1$.

Uncertainty

Models that contain uncertainty lead to expectational Euler equations. Add to the discrete-time example that the agent believes income y_s for $s > t$ to be stochastic from the perspective of period t. The Euler equation becomes

$$u'(c_t) = \beta R\hat{E}[u'(c_{t+1})|I_t] \tag{5}$$

where $\hat{E}[.|I_t]$ represents the agent's expectation given information set I_t. The stochastic version of the consumption Euler equation has an analogous interpretation to that under certainty: the household equates *expected* (discounted) marginal utility over time.

Taking a second-order approximation to marginal utility in $t + 1$ around c_t and re-organizing gives

$$\hat{E}\left[\frac{c_{t+1} - c_t}{c_t}|I_t\right] = \sigma_t\left(1 - (\beta R)^{-1}\right) + \frac{1}{2}\varphi_t E\left[(c_{t+1} - c_t)^2|I_t\right]$$

where $\varphi_t = -\frac{c_t u'''(c_t)}{u''(c_t)}$ is the coefficient of relative prudence (see for example Dynan 1993). It is now *expected* consumption growth that rises with the real interest rate and falls with impatience. Additionally, for $\varphi_t > 0$, risk leads to precautionary saving: higher expected consumption growth (much like liquidity constraints).Finally, actual consumption growth is also driven by the realization of uncertainty about current and future income.

Testing and Estimation

An expectational Euler equation is a powerful tool for testing and estimating economic models in large samples, because, along with a model of expectations, it provides orthogonality conditions on which estimation can be based. Only randomization, as under experimental settings, delivers such a clean basis for estimation without near-complete specification of an economic model, including the sources of uncertainty.

Considering our main example, define $\varepsilon_{t+1} = u'(c_{t+1}) - (\beta R)^{-1}u'(c_t)$. Hall (1978) pointed out that Eq. (5) implies that $\hat{E}[\varepsilon_{t+1}z_t|I_t] = \hat{E}[\varepsilon_{t+1}|I_t] = 0$ for *any* z_t in the agent's information set, I_t. Under the assumption of rational expectations, mathematical expectations can be used in place of the agent's expectations. Thus, this equation predicts that observed changes in discounted marginal utility are unpredictable using I_t, or that marginal utility is a martingale, a strong theoretical prediction that Hall (1978) tests. Hansen and Singleton (1983) use a version of the stochastic Euler equation with a portfolio choice as the basis for estimation (and testing) of the

parameters of the representative agent's parameterized utility function.

Following these papers (and others), large-sample testing and estimation of Euler equations under the assumption of rational expectations has played a central role in the evaluation of dynamic economic models. Most research applies the generalized method of moments (GMM) of Hansen (1982) using the restrictions on the moments of time series implied by the expectational Euler equation. Considering a $J \times 1$ vector of z_t's, $\mathbf{z}_t$, and, based on our example, define the column vector $\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t) = (\beta R u'(c_{t+1}) - u'(c_t))\mathbf{z}_t$, so that we have the J moment restrictions $E[\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t)] = 0_{J\times 1}$. For example, letting $u'(c_t) = c_t^{-(1/\sigma)}$ and assuming that second moments exist and the model is covariance stationary, the time-series average of $\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t)$ should converge to $E[\mathbf{g}(c_{t+1}, c_t, \mathbf{z}_t)]$ for the true σ, β and R. The GMM estimates of σ, β and R are those that minimize the difference (according to a given metric) between the observed empirical moments and their theoretical counterparts, $0_{J\times 1}$.

This general approach has the advantage that complete specification of the model is not necessary. In our example, the stochastic process for income need not be specified nor the stochastic process for consumption determined (which can be quite demanding in terms of computer programming and run-time). That said, more complete specification can give more theoretical restrictions and thus more power in asymptotic estimation. Gourinchas and Parker (2002), for example, uses numerical methods to bring more theoretical structure to bear in estimation. Further, more complete specification can allow one to use small-sample distribution theory and thus avoid the approximations inherent in using asymptotic distribution theory for inference in finite samples. A recent cautionary example is provided by the literature showing that standard asymptotic inference can be highly misleading in large samples with 'weak instruments'.

See Also

- ▶ Bayesian Methods in Macroeconometrics
- ▶ Calculus of Variations
- ▶ Dynamic Programming
- ▶ Generalized Method of Moments Estimation
- ▶ Instrumental Variables
- ▶ Liquidity Constraints
- ▶ Permanent-Income Hypothesis
- ▶ Precautionary Saving and Precautionary Wealth
- ▶ Ramsey Model
- ▶ Rational Expectations Models, Estimation of

Bibliography

Dynan, K. 1993. How prudent are consumers? *Journal of Political Economy* 101: 1104–1113.

Gourinchas, P.O., and J.A. Parker. 2002. Consumption over the life cycle. *Econometrica* 70: 47–89.

Hall, R.E. 1978. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy* 86: 971–987.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054.

Hansen, L.P., and J.K. 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. Synamic choices of hyperbolic consumers. *Econometrica* 69(4): 935–958.

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

Tintner, G. 1937. Monopoly over time. *Econometrica* 5: 160–170.

Weitzman, M.L. 2003. *Income, wealth and the maximum principle*. Cambridge, MA: Harvard University Press.

Euler's Theorem

Peter Newman

Keywords

Adding-up problem; Calculus of variations; Euler equations; Euler's Theorem; Homogeneous functions; Lagrange multipliers

JEL Classifications

C6

Euler's Theorem on homogeneous functions is one of those useful pieces of multivariable calculus that has tended not to receive the attention in mathematical textbooks that its importance in economic theory warrants. An analogous case is Lagrange multipliers, though there the analysis in most textbooks falls far short of the rigour and depth that are needed for fruitful economic applications, as it often does of Euler's other discovery of direct importance in economics, the so-called Euler equations in the calculus of variations (for a critical discussion, see Young 1969). With Euler's Theorem there are no such worries, however, and the discussion in a work like that of Courant (1936, vol. 2, pp. 108–10) is quite adequate.

Once the necessary notation and terminology is established, the statement of the theorem follows easily. Given any real number k, let F be a real-valued function defined on some non-empty subset S of vectors $x \in R^n$.Then F is said to be *homogeneous of degree k* (*h.d.k.*) if the equation

$$F(tx) = t^k F(x) \tag{1}$$

holds for every x $\in S$ and every real number t.

Let f be a differentiable function defined on a non-empty open subset $G \subset R^n$, and denote the *gradient* of f at x, that is, the n-dimensional vector of its partial derivatives f_i evaluated at x, by $\nabla f(x)$. The inner product of any two vectors a and b is written $\langle a,b \rangle$.

Euler's Theorem The differentiable function f is homogeneous of degree k if and only if the following Euler relation holds for every $x \epsilon G$,

$$\langle x, \nabla f(x) \rangle = kf(x) \tag{2}$$

For a proof, see Courant (1936). Notice that this theorem *characterizes* homogeneous functions, that is, any function satisfying (2) for all x must satisfy (1), hence be *h.d.k.* A simple but often useful corollary of the theorem is that, if f is r-times differentiable and $m \leq r$, then each of its partial derivatives of order m is homogeneous of degree $k - m$, so that each f_i is *h.d.*$(k-1)$, each f_{ij} is *h.d.*$(k-2)$, and so on. Since homogeneous functions crop up almost everywhere in economics, Euler's Theorem is a standard tool with innumerable applications. So it is slightly odd that what was apparently its first use occurred so late, and that it was not by an established mathematical economist. In his review of Wicksteed (1894), A.W. Flux (1894) pointed out that Wicksteed could have saved himself a great deal of trouble if he had simply cited Euler's Theorem instead of, in essence, proving it all over again. It was indeed in the controversy over the so-called adding-up problem in the theory of distribution that Euler's Theorem first gained notoriety. For details of the adding- up problem, see Steedman (1987); here only a few of the main points will be lightly sketched in.

Assume that the firm wishes to minimize the cost of producing a scalar output s by the use of factors $x_1, x_2, \ldots, x_n = x$ bought at competitive prices $p_1, p_2, \ldots, p_n = p$. Under standard assumptions the first order conditions for this minimization yield

$$p_i = \lambda f_i(x) \tag{3}$$

where $i = 1, 2, \ldots, n$ and λ is the associated Lagrange multiplier. This multiplier is of course the marginal cost of output, a fact which can be guessed at from (3) on purely dimensional grounds alone. Assume now that the production function f, where $\eta = f(x)$, is homogeneous of some unknown degree k.

Then, substituting from (3) into (2) and remembering the meaning of $\nabla f(x)$,

$$\lambda^{-1} \langle x, p \rangle = kf(x). \tag{4}$$

If there is competition in the product market as well, that is, free entry, then in long-run equilibrium marginal cost will equal the price q of the product, so that (4) becomes

$$\langle x, p \rangle = kqf(x). \tag{5}$$

The left-hand side of (5) is total factor payments. If constant returns prevail f is *h.d.*1 and so $k = 1$. All is well, since the right-hand side is then total revenue, equal to the sum of factor payments. If on the other hand $k < 1$, so that returns to scale

are decreasing, (5) shows that there will be something left over after all the purchased inputs have been paid.

What this residual really means is not clear. Some writers have interpreted it to be the returns (rent) to some non-marketed factor internal to the firm. But in that case why isn't the factor sold by its owner to the firm (after all, we are in long-run equilibrium, so that quasi-rents do not apply)? Or is there no external market for the factor?

This is not the place to go into such qsts, but it may be suggested that the incompleteness resides more in the theory than in either the markets or the factor payments. If $k > 1$ there are increasing returns to scale, and (5) then suggests that there will not be enough revenue to meet total factor payments. But with increasing returns the hypothesis of perfect competition in the product market has to be abandoned, so the passage from (4) to (5) is illegitimate and (5) does not hold.

Bibliography

Courant, R. 1936. *Differential and integral calculus*. 2 vols. London: Blackie & Son.

Flux, A.W. 1894. Review of Wicksteed (1894). *Economic Journal* 4: 305–308.

Steedman, I. 1987. Adding-up problem. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. London: Macmillan.

Wicksteed, P.H. 1894. *An essay on the co-ordination of the laws of distribution*. London: Macmillan.

Young, L.C. 1969. *Lectures on the calculus of variations and optimal control theory.* Philadelphia: W.B. Saunders Co..

Euro

Adam S. Posen

Abstract

The euro has been a limited success at home, but it has not challenged the dollar as a global reserve currency. This reflects the limited impact of the euro on member economies. While European financial markets and trade are far more integrated than before adoption of the euro, that is the result of broader international trends. Factors inhibiting growth in the eurozone's real economy prevent truly deep financial integration as well, despite the removal of currency risk. The euro has delivered price stability and credible monetary policy, but not induced the convergence and reform necessary to improve European economic performance.

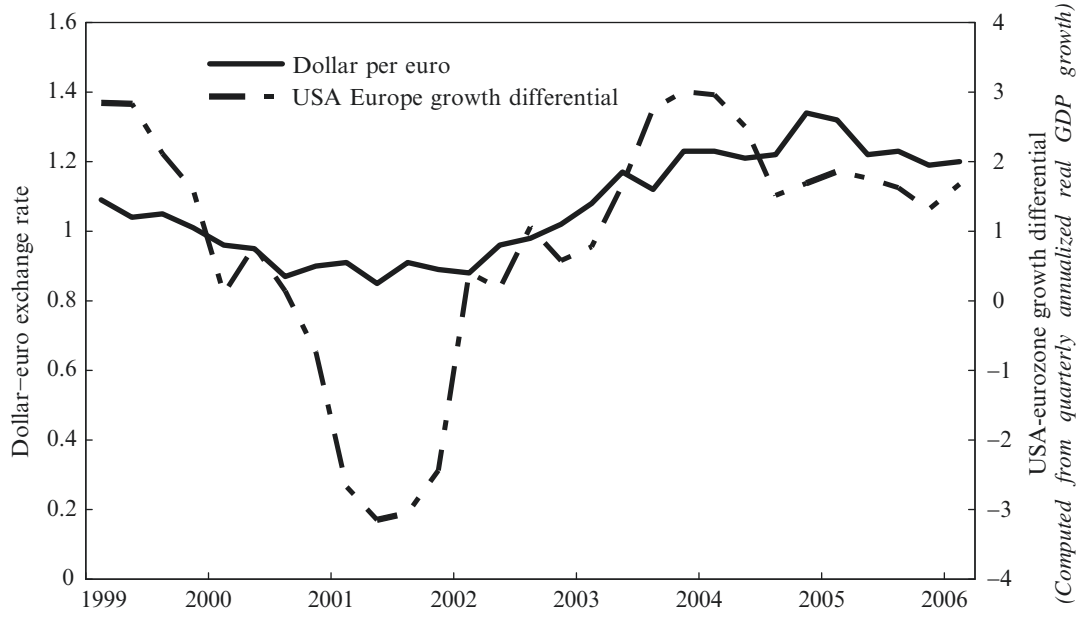
Keywords

Currency unions; Euro; European Central Bank; European Monetary Union; Eurozone; Exchange Rate Mechanism (EU); Exchange rate uncertainty; Financial market integration; Inflation; Inflationary expectations; International capital flows; Money market integration; Portfolio investment; Stability and Growth Pact (EU)

JEL Classification

F

The launch of the euro, the European Union's currency (at least for 12 of the 27 current members), on 1 January 1999, was a birth long foretold. From at least the 1992 Maastricht Treaty onwards, its creation was at the forefront of the European overall integration agenda, and the meeting of criteria for eurozone entry dominated macroeconomic policymaking in Western Europe. The academic and policy discussion of European Monetary Union's (EMU) potential advantages and disadvantages began even earlier (see Canzoneri et al. 1992; De Cecco and Giovannini 1989; De Grauwe 2000; Cecchini 1988; as well as the seminal European Commission 1990. Most of these studies concerned how best to make EMU work, taking the goal as a given, or assessing the optimality of the EU as a currency area.) New international reserve currencies, as the euro has begun to be, do not come along every day, or even every century. New currencies in general are launched usually out of



Euro, Fig. 1 Dollar-euro exchange rate and real GDP growth differential, 1999–2006 (Source: IMF, IFS Statistics)

need, due to replacement of a currency of hyperinflation-eroded value or to political fragmentation or secession; when currency unions are formed, they are usually done as pegs to a previously existing anchor currency of the largest and/or most stable member economy. The voluntary adoption of the euro by sovereign but not politically unified nations, and its replacement of already stable currencies (notably the Deutschmark), is thus an extraordinary monetary experiment and policy undertaking.

While the euro certainly has had no shortage of champions among economists – including beyond Euroland's borders the economists Bergsten (1997), Eichengreen (1999), Mundell (1998), and Portes and Alogoskoufis (1991) – many monetary economists observing the euro have tended to be sceptical: first of the virtues of the goal of monetary integration in Europe itself, then of the project's political viability, and then of its economic sustainability, in turn asserting that the euro was a solely political project. (Notable examples of this scepticism include, on the political side, Currie et al. 1992; Walters 1990; and, famously, Feldstein 1997; and on the economic side Arestis and Sawyer 2001; De Grauwe 1996; Dornbusch 1989; Giavazzi and Spaventa 1990; and Weber 1991. See also the essays by eurosceptics in the face of mounting contrary evidence collected in *Cato Journal* 2004.) Only as the euro passed its eighth birthday in wide usage, remained well past parity with the US dollar (see Fig. 1) and experienced a strong cyclical recovery in the eurozone has sentiment changed. Increasingly, the question is being raised whether the euro might appreciate against the dollar for an extended period, be the beneficiary of substantial international portfolio adjustments, or even begin to supplant the dollar as the dominant global reserve currency. (Recent examples include Chinn and Frankel (2004, 2007); Obstfeld and Rogoff 2004; and Summers 2004.) The euro's viability in its own large economic area may not be sufficient to set it on a path to monetary leadership, but its existence now presents an alternative for capital markets to turn to should the dollar's own appeal diminish.

The waiting for US missteps for the euro to rise in importance, however, is a critical commentary on the limitations of the euro's importance to the eurozone member economies' performance in and of itself. When the euro was first proposed, a number of studies claimed that monetary

integration would bring significant direct benefits to the economic performance of member states. Emerson et al. (1992) estimated that the elimination of transaction costs from moving to a single European currency would yield direct benefits of up to 0.4 % of EU GDP; the European Commission (1996) estimated cost savings of 1.0 % of GDP simply from eliminating transaction costs. The European Commission (1990) made the case that the reduction of nominal and real exchange rate uncertainty would lead to significant growth in intra-EU trade and investment. Financial markets in particular were expected to benefit from the introduction of the euro – McCauley and White (1997) and the European Commission (1997) forecast a rapid deepening and liquidity increase in European bond and lending markets, and perhaps even a 'decoupling' of European interest rates from those of the United States.

While empirical investigations to date of these effects remain mixed in interpretation, there is no question that the real economic effects of the euro's launch on the eurozone member countries have been something of a disappointment. In particular, European financial markets and trade integration are far deeper today than they were before the adoption of the euro, yet how much this represents the effect of the euro on EU integration as opposed to the broader international trends towards global integraton that benefited non-euro members as well is in doubt (see Forbes 2005; Lane and Wälti 2006; Mann and Meade 2002; and Rey 2005). The eurozone's interest rates remain asymmetrically affected by US interest rates, at least through the early 2000s, as established by Chinn and Frankel (2004). The effect of the euro on price convergence and on macroeconomic discipline cannot be all that substantial if on net there has been limited visible improvement in either of these areas (see the assessments of price convergence in Bradford and Lawrence (2003, 2004); and Rogers 2003 and of macroeconomic discipline in Posen (2005a, b). It seems that the euro has proven on net 'irrelevant' to real growth performance of large Continental European economies, neither a harm nor a boon to them, as Posen (1998) forecast it would be.

The External Opportunities and Shortfalls for the Euro

The degree to which the euro comes into wider usage beyond intra-eurozone transactions, for example as an invoicing currency in world trade, is a major issue because of the eurozone's already large share of world output and trade (roughly comparable to that of the United States) and of the established 'domestic' monetary stability of the eurozone. Size does matter for international currency purposes. Yet insufficient integration and depth of European financial markets as well as lagging economic performance remain constraints on the euro's wider adoption and usage. Also important is the lack of coherent institutional representation for the eurozone in international monetary forums. Compared with the EU's one voice in global trade negotiations, the inability of the eurozone to speak as a single entity is striking, especially given the unconsolidated overrepresentation of the eurozone in the Bretton Woods institutions.

History also plays a role, however, in the global demand for currencies and their strength. Inertia and incumbency clearly contributed to the lingering of the British pound in a significant share of international reserves well after the Second World War. Yet the combination of macroeconomic mismanagement and growth underperformance in the United Kingdom from the 1920s to the 1980s eroded that role, and it is worth remembering that the passing of international monetary leadership from the pound to the dollar in the mid-twentieth century was in large part driven by these factors undermining the pound's reserve status. The steady accumulation of international debt by the United States since 1991 could contribute to a similar switch now that the euro is available. An extended dollar depreciation, the natural reaction to a multi-year series of widening US current account deficits, could induce a persistent portfolio diversification into euros by private and official holders of dollars.

In Washington, Frankfurt and Brussels, however, the widespread governmental opinion remains that the euro will not close the gap in usage with the dollar until the eurozone closes the gap with the US economy in per capita GDP

growth and employment on a sustained basis (Fig. 1 shows the growth differential of the USA over the eurozone). In a typical official expression of this sentiment, Quarles (2005, 40) finds that 'too much attention is being focused on exchange rate[s]... and too little on what seems... of far greater importance: namely, the more effective functioning of economies with regards to growth in output and employment'. Successive US governments have viewed both the short-term international adjustment process and the longer-term role of the euro vis-à-vis the dollar as driven by the gap in growth rates between the USA and Europe – with the burden on European economies to catch up by raising their growth rates. EU officials' disappointment with the degree of structural reforms catalysed by the introduction of the euro echoes this view, as does the promotion of the Lisbon Agenda announced in March 2000 for promotion of growth in the EU.

Such an external relative focus overlooks one achievement of the launching of the euro – ending the succession of devaluations, competitive depreciations and currency crises that had beset the members of the Exchange Rate Mechanism (ERM) prior to 1999. Certainly, the experiences of intra-European depreciations upon countries leaving the ERM, especially those of 1992–3, and their impact on economic performance and political outcomes in member states were in the forefront of European policymakers' minds when the run-up to the euro was under way in the late 1990s. And, despite the divergence in histories of some eurozone members, inflation and inflation expectations have remained stable and low in the eurozone. That could have been expected to assist in trade promotion among the already interdependent eurozone economies (see currency unions).

Still there has been little or no expansion in trade as a result of the adoption of the euro – among other evidence, the share of total eurozone exports destined for other members of the eurozone did not increase with the introduction of the currency, as would have been likely if the common currency had promoted trade (Baldwin 2005 provides an excellent analytical summary of the evidence on this score). As shown in Rogers (2003), the bulk of convergence in traded-goods prices within the eurozone occurred between 1990 and 1994, in response to the creation of the single market, and not after 1999 and the introduction of the euro. As for the global dimension, there has been little change in the share of foreign exchange transactions denominated in euros globally from that previously denominated in Deutschmarks. Similarly, the use of the euro as an invoicing currency is somewhat higher than that for the eurozone home currencies prior to EMU, but remains far from universal within Europe or even comparable to the dollar's usage (with the regional exception of some of the newest members of the European Union).

Even the spreading use of the euro in the EU's new members in the east has been far less than many might have expected. A critical part of this outcome has been the insistence on the part of the European Central Bank (ECB) that all prospective eurozone members go through the full Maastricht Treaty-specified process for qualification, including not just fiscal discipline and nominal convergence but also a two-year period in the 'waiting room' of a new ERM-II mechanism. Early, expedited or unilateral adoption of the euro in EU member countries has in fact been discouraged by the ECB (with the exception of Estonia's pre-existing currency board with the euro). Arguably, this has as much to do with the ECB's desire for perceived control over monetary developments, given the ECB's Bundesbank-esque 'two pillar' strategy (of looking at both monetary growth and inflation goals when setting policy), and for keeping decision-making in the ESCB manageable, as with maintaining necessary discipline on eurozone members (see European Central Bank). The ECB has also been explicitly opposed to 'euroization' (dollarization with euros) by non-EU member countries, again partly for monetary control reasons, albeit acknowledging its contribution to stability in the post-conflict Balkan economies.

The Limited Impact of the Euro on the Eurozone Financial Integration and Performance

The euro has delivered monetary stability in the face of a long list of economic shocks and a large

initial decline against the dollar, only to rebound strongly since Autumn 2001 (see Fig. 1). Europe has failed to follow the creation of the euro with the complementary policy reforms that were widely expected, however. This leaves an underlying tension between the constraints on national economic policy measures such as those in the Stability and Growth Pact on fiscal policy (see stability and growth pact) and the national frustrations with poor economic performance – a tension that raises recurrent doubts in eurosceptic financial markets about the sustainability of the euro itself, despite its lack of obvious vulnerabilities or viable exit options for any member country.

The euro was widely expected to transform two aspects of the eurozone economies: the integration and depth of their financial markets, and the conduct of their macroeconomic policies. Particularly with regard to the former, there has been beneficial change at least partly attributable to the euro's introduction and acceptance. Money market integration, which is critical to the implementation of a single monetary policy for the eurozone, given the need to transmit monetary policy in a decentralized fashion across the member economies, has succeeded. It took European money markets less than a month in 1999 to 'learn' how the new operational framework functioned, and to eliminate most of the volatility and cross-border dispersion in overnight interest rates. The evidence of integration in the unsecured lending rates in the European money market is similarly clear. Rey (2005) finds that government bond markets have seen intra-eurozone interest rate spreads virtually disappear, and benchmark securities of different countries have begun to emerge. Corporate bond markets went from 'almost non-existent' prior to EMU to 150 billion euro of issuance in 2003, and the euro swap market has become the largest financial market in the world.

Eurozone financial markets, however, still have a long way to go to become a global competitor with those based in London or New York. Factors in the non-financial economy, such as legal differences, obstacles to more rapid real growth, transaction costs, and institutional gaps in financial supervision combine to keep the eurozone from achieving truly deep, integrated financial markets, despite the removal of currency risk. Thus, there remains a striking contrast between the repo (repurchase of safe assets at central banks) and unsecured market in the degree of cross-national differences in interest rates due to the ongoing lack of harmonization in legal and procedural treatment of financial instruments in the eurozone countries. The costs of making cross-border securities transfers within the eurozone can still be ten times more than the cost of securities transfers within a given eurozone country.

Given the surge in capital flows across borders worldwide, following the recovery from the 1997–98 Asian financial crisis, almost half of which were in the form of portfolio investment, one would expect greater influence of market opinion about assets in a given currency or region upon the actual allocation of capital between regions. It seems that prospects for economic growth drive the relative demand for a region's assets, mostly by determining where trade and investment expands, which then in turn sets the pace of stock market integration of that region with the rest of the world. Given the medium-term outlook for European growth, this appears to militate against an increase in investment and therefore in integration (and influence) of European capital markets, which might be partially offset by some diversification incentives. In the long run, though, a slow growth rate in Europe would also translate into a smaller share of global GDP, and less incentive for central banks to hold euro-denominated reserves. In this context, Forbes (2005) and Lane and Wälti (2006) independently investigate whether the euro's launch prompted greater co-movement of stock prices within the eurozone across national borders, indicating greater financial integration as a result of EMU. Both investigations find that stock market correlations of eurozone member markets with the United States increased after the introduction of the euro more than those between the eurozone countries.

Prospects for the Euro

The euro therefore occupies something of a halfway house. In terms of its purely technical

functions it has been a resounding success, with no problems in acceptance at home or abroad, or in the payments system, and there has been convergence in key eurozone money market interest rates. There has also been evidence of stable low-inflation expectations for the varied eurozone membership as a whole, which remains an outstanding achievement of European central banking. None of the broader forecasts of economic doom or internal political conflict predicted by (mostly American) Chicken Littles came to pass, and those predictions look less credible than they ever did. European financial markets have significantly deepened and added liquidity since the advent of the euro, particularly for fixed-income securities. The sheer size of the eurozone economy as well as the ongoing adjustment of the world economy to US current account deficits propel the euro towards a prominent global role.

At the same time, however, European relative economic performance and growth potential will continue to fall short of that of many other advanced economies and large emerging markets for the foreseeable future. The adoption of the euro and the associated convergence process have failed to induce, let alone produce, the needed transformation in European economic structures, policies and performance. In most scenarios, a collapse of the dollar in coming years, or even an ongoing orderly adjustment involving higher US long-term interest rates and lower net imports, will have at least as great a contractionary effect on the eurozone as it will on the US economy – even if the Asian currencies take on their share of the adjustment burden. And if the Asian currencies, notably the Chinese yuan and Japanese yen, play their part, reserve switches accruing to euro-denominated securities, and their political benefits, will diminish along with the euro's share in the adjustment process. And as yet there has been little evidence of a change in global invoicing patterns from dollars to euros for traded good transactions.

In short, the euro has been a success within limits at home, but the eurozone economy is not yet strong enough – and is unlikely to be so for some time – to challenge the dollar as a global reserve currency or even to be widely utilized outside its borders. The euro, however, is not judged solely on its own merits, either by markets or by the international community, but rather is judged also in relative terms against developments in the dollar zone and elsewhere.

See Also

- ▶ Currency Unions
- ▶ European Central Bank
- ▶ European Monetary Union
- ▶ Stability and Growth Pact

Bibliography

Arestis, P., and M. Sawyer. 2001. Will the euro bring economic crisis to Europe? Economics working paper archive: Washington University in St Louis.

Baldwin, R. 2005. The euro's trade effects. Presented at the ECB workshop: What effects is EMU having on the euro area and its member countries? Presented at the European Central Bank workshop. Frankfurt, 16 June.

Bergsten, F.C. 1997. The impact of the euro on exchange rates and international policy cooperation. In *EMU and the international monetary system*, ed. M. Paul, T.H. Krueger, and B. Turtelboom. Washington, DC: International Monetary Fund.

Bradford, S., and R.Z. Lawrence. 2004. *Has globalization gone far enough? The costs of fragmented international markets*. Washington, DC: Institute for International Economics.

Canzoneri, M.B., V. Grilli, and P. Masson. 1992. *Establishing a Central Bank for Europe*. Cambridge: Cambridge University Press.

Cato Journal. 2004. *The future of the euro*, vol. 24 (Special Issue). Washington, DC: Cato Institute.

Cecchini, P. 1988. *Research on the 'Cost of Non-Europe': Basic findings* (Cecchini report). Luxembourg: EUR-OP.

Chinn, M.D., and J.A. Frankel. 2004. The euro area and world interest rates. Working paper series 1016. Center for International Economics, UC Santa Cruz.

Chinn, M.D., and J.A. Frankel. 2007. Will the euro eventually surpass the dollar as leading international reserve currency? In *G7 current account imbalances: Sustainability and adjustment*, ed. R. Clarida. Chicago: University of Chicago Press.

Currie, D., P. Levine, and J. Pearlman. 1992. European monetary union or hard EMS? *European Economic Review* 36: 1185–1204.

De Cecco, M., and A. Giovannini. 1989. *A European Central Bank? Perspectives on monetary unification after ten years of the EMS*. Cambridge: Cambridge University Press.

De Grauwe, P. 1996. Monetary union and convergence economics. *European Economic Review* 40: 1091–1101.
De Grauwe, P. 2000. *Economics of monetary union*. Oxford: Oxford University Press.
Dornbusch, R. 1989. The dollar in the 1990s: Competitiveness and the challenges of new economic blocs. In *Monetary policy in the 1990s: A symposium*, sponsored by the Federal Reserve Bank of Kansas City, Jackson Hole, WY.
Eichengreen, B. 1999. Will EMU work? In *Euroen og den norske kronens skjebne*, ed. A.J. Isachsen and O.B. Roste. Bergen: Fagbokforlaget. English version online. Available at http://emlab.berkeley.edu/users/eichengr/policy/merrill.pdf, accessed 25 February 2007.
Emerson, M., D. Gros, and A. Italianer. 1992. *One market, one money: An evaluation of the potential benefits and costs of forming an economic and monetary union*. Oxford: Oxford University Press.
European Commission. 1990. One market, one money. *European Economy* 44.
European Commission. 1996. *Economic evaluation of the internal market*. European economy: Reports and studies no. 4. Brussels.
European Commission. 1997. External aspects of economic and monetary union. Euro papers No. 1. Brussels.
Feldstein, M. 1997. The political economy of the European economic and monetary union: Political sources of an economic liability. *Journal of Economic Perspectives* 11(4): 23–42.
Forbes, K. 2005. The euro and financial markets. In Posen (2005a).
Giavazzi, F., and L. Spaventa. 1990. The 'new' EMS. Discussion paper No. 369, CEPR.
Lane, P., and S. Wälti. 2006. *The euro and financial integration*. Discussion paper. Dublin: Institute for International Integration Studies.
Mann, C.L., and E.E. Meade. 2002. *Home bias, transaction costs and prospects for the euro: A more detailed analysis*. Working paper. Washington, DC: Institute for International Economics.
McCauley, R.N., and W.R. White. 1997. *The euro and European financial markets*. Working paper No. 41. Basle: BIS.
Mundell, R. 1998. What the euro means for the dollar and the international monetary system. *Atlantic Economic Journal* 26: 227–237.
Obstfeld, M., and K. Rogoff. 2004. *The unsustainable US current account position revisited*. Working paper No. 10869. Cambridge, MA: NBER.
Portes, R., and G. Alogoskoufis. 1991. International costs and benefits from EMU. *In The economics of EMU. European Economy* 1 (Special Issue), 231–245.
Posen, A.S. 1998. Why EMU is irrelevant for the German economy. Working paper No. 1998/11, Center for Financial Studies, University of Frankfurt.
Posen, A.S. (ed.). 2005a. *The euro at five: Ready for a global role?* Washington, DC: Institute for International Economics.
Posen, A.S. 2005b. Can Rubinomics work in the eurozone? In Posen (2005a).
Quarles, R. 2005. *Discussion of the euro and the dollar*. In Posen (2005a).
Rey, H. 2005. *The euro and financial markets*. In Posen (2005a).
Rogers, J.H. 2003. Monetary union, price level convergence, and inflation: How close is Europe to the United States? International Finance discussion papers No. 740. Washington, DC: Board of Governors of the Federal Reserve System.
Summers, L.H. 2004. *The United States and the global adjustment process*. Third Annual Stavros S. Niarchos Lecture. Washington, DC: Institute for International Economics. Online. Available at https://piie.com/commentary/speeches-papers/united-states-and-global-adjustment-process ResearchID = 200. Accessed 25 Feb 2007.
Walters, A. 1990. *Monetary constitutions for Europe*. Speech at the 28th meeting of the Mont Pèlerin Society, Munich.
Weber, A.A. 1991. EMU and asymmetries and adjustment problems in the EMS: Some empirical evidence. *European Economy* 44: 187–207.

Euro Zone Crisis 2010

Daniel Gros and Cinzia Alcidi

Abstract

The euro zone crisis is commonly regarded as a sovereign debt crisis. This definition certainly applies to Greece, but the Irish case represents an almost pure specimen of a banking crisis voluntarily transformed into a sovereign crisis. A debt crisis in two small, peripheral economies could become systemic because the financial system of the euro area is overstretched and highly integrated. Had the Greek and Irish crises occurred when euro zone banks were strong and/or not very interconnected, the euro zone crisis would not have happened.

Keywords

Euro zone; Bailout; Banking crisis; Leverage; Sovereign debt

JEL Classifications
E60

Introduction

The euro zone crisis started in early 2010 when it emerged that the Greek government had for years doctored the official data on its deficits and debt. The figures for the deficit and debt level presented by the new government were so much higher than the previous ones that rating agencies and many market participants downgraded their assessment of Greece's ability to service its debt fully. As a result, the cost of refinancing the Greek debt increased sharply and the government could not secure the resources needed to fund its current deficit and roll over the portion of the debt coming due. By the end of April 2010 it had to be bailed out with a €110 billion programme.

The second stage of the crisis came about six months later when it emerged that the Irish government had been 'misled' about the scale of the losses in its banks. As the Irish government had guaranteed all the liabilities of its banks it was now itself on the brink of insolvency. Moreover (although this was not made public at the time), the ECB had become uncomfortable with the huge exposure it had to Irish banks, which had become totally dependent on central bank financing. The ECB therefore pushed the Irish government to recapitalize its banks, but this could be done only with outside help. The Irish government had thus little choice but to apply for external financial support.

With the Greek and Irish bailouts, the euro zone has shown the world two pure specimens of financial crisis: one originated by the mismanagement of fiscal policy (Greece), the other by mismanagement of a credit bubble and banking supervision (Ireland). The Portuguese crisis, which emerged in early 2011, seems to represent a hybrid specimen: a combination of a fiscal crisis (like Greece) and a private debt crisis (like Ireland).

A Brief Chronology

Although Greece accounts for a small portion (less than 3%) of the euro area GDP (and even less of its banking assets), in early 2010 financial markets reacted strongly to the prospect of a sovereign insolvency. A first consequence of the realization that Greece would not be solvent without external financial support was that investors started to price more widely government solvency in the bond market. As a result, the risk premia on the debt of other countries with weak fundamentals also rose. But more important was a generalized increase in risk aversion, which led to a fall in the prices of all risky assets in a similar vein (but of course a much less severe magnitude) as after the collapse of Lehman Brothers in late 2008.

The European banking sector was particularly affected because it was widely believed that a number of banks would not survive a default by Greece. However, which banks held how much of Greek debt was not known. In an environment of widespread risk aversion and many highly leveraged banks this resulted in a drying up of parts of the interbank market, which performs a vital role in the financial system.

The German government reiterated on several occasions its aversion to a bailout, stressing that this must be only an *ultima ratio* mechanism. But when faced with the spectre of a 'second Lehman crisis' and the prospect of large losses in the weak German banks heavily exposed to Greece and other peripheral countries, it had no choice but agree to a rescue package of about h110 billion. This is an EU/IMF rescue package according to which the IMF provides support under a three-year €30 billion standby arrangement (the IMF's standard lending instrument) while euro area members pledge a total of €80 billion in bilateral loans against the implementation of strict austerity measures monitored by the IMF. The sum agreed is supposed to fully finance Greece's remaining deficits (and rollover obligations) during the following three years. It was assumed then (on the basis of experience with 'normal' IMF programs) that Greece would be able to access private capital markets at reasonable rates towards the end of this period. However, in early 2011 it became clear that the hypothesis was far too optimistic. In March the terms and the conditions of the loans to Greece were reviewed to include an extension of the maturity and lower interest rates.

In the spring of 2010, Europe's leaders also thought that Greece was a unique and special case and that no other country would ever need financial support. However, only a few days after the Greek rescue, financial markets went into such a tailspin (risk premia rose, some markets ceased to function) that a new and much larger financing mechanism had to be hastily created.

During the dramatic weekend of 9 May 2010, two financing mechanisms were set up in order to allow the authorities to react to future financial crises in a more coordinated and organized manner. The headline figure of the total potential funding was €750 billion, to be provided by three different entities: €60 billion, guaranteed by the EU budget, coming through a newly created European Financial Stabilization Mechanism (EFSM); €440 billion, guaranteed on a pro rata basis by euro area member states, coming through the also newly created European Financial Stability Facility (EFSF); and up to €250 billion from the IMF.

Together with the ECB interventions in the euro area public and private debt securities markets (Securities Markets Programme) aiming at ensuring liquidity in those market segments judged to be 'dysfunctional', this package did restore stability in the financial markets for a few weeks.

In early June 2010, since tensions in the interbank market persisted, member states and the European Institutions (Commission and Committee of European Banking Supervisors, CEBS) agreed to make public for the first time the results of ongoing stress tests for major European banks.[1]

The rationale for the tests was to disclose information about the state of the European banking system in order to dissipate doubts about their resilience. The Spanish supervisory authorities were particularly keen on this move because they hoped that by showing that their banks were 'safe and sound', it would be easier for Spanish banks to regain access to the interbank market. More generally, the publication of the stress tests was supposed to prove that the most important banks had sufficient capital to withstand even a so-called 'adverse' scenario. This should have improved confidence in the banking system in general.

Yet the objective of the exercise was achieved only temporarily.[2] During the summer of 2010 risk premia on the government bonds of the four 'fiscally challenged' countries (Portugal, Ireland, Greece and Spain) started to increase again. This accelerated after a Franco-German agreement in Deauville on economic governance and the decision by the European Council of 28 October to establish a permanent crisis mechanism to safeguard the financial stability of the euro area. This decision proved to be a watershed because it suggested a change in the ground rules of peripheral euro area debt markets: on that occasion all 27 Member States agreed on the proposal (then submitted to the European Council and implemented in early 2011) for a limited, technical Treaty amendment to provide a legal basis for establishing a permanent crisis mechanism. In March 2011, the European Council adopted the basic features of the new device: the European Stability Mechanism (ESM). The ESM, which will be operational as of mid-2013, is based on the existing EFSF but, unlike the EFSF, the provision of liquidity is conditional to a debt sustainability assessment (conducted by the European Commission and the IMF, in liaison with the ECB). In the event that the analysis reveals that a member state is insolvent, the country is expected to negotiate a comprehensive plan with its private creditors. Moreover collective action clauses (CACs) will be included in the terms and conditions of all new euro area sovereign bonds, starting in June 2013. These clauses should provide the legal basis for the negotiation process with creditors and enable them to pass by qualified majority a decision agreeing a legally binding change to the terms of payment. This could take different forms (standstill, extension of maturity,

[1] http://stress-test.c-ebs.org/documents/Summaryreport.pdf.

[2] See, among others, Veron (2010) and Blundell-Wignall and Slovik (2010).

interest-rate cut and/or haircut) depending on the specific case, but clearly implies that if losses materialize they will be borne, at least partially, by the private sector.

Financial markets did not welcome this approach, and Ireland became the first victim of deteriorating market conditions. Indeed, market pressures on Ireland had started mounting in October 2010, when the Irish government decided to rescue some of its banks that had published losses that were considerably higher than estimated a few months earlier. The high costs of this bank bailout program resulted in a deficit of 32% of GDP, and the risk premia on Irish government (and bank) bonds shot up. As a consequence the Irish government quickly had to ask for external support. On 28 November, an h85bn financial assistance package was agreed and Ireland committed to a sweeping restructuring of its banking system and even more sweeping budget cuts. According to the rescue plan, the EU provides financial assistance for €45bn, through the European Financial Stability Mechanism and the European Financial Stability Fund, together with bilateral loans from the UK (€3.8bn), Sweden (€0.6bn) and Denmark (€0.4bn). The IMF provides h22.5bn and the Irish sovereign €17.5bn through the Treasury cash buffer and investments of the National Pension Reserve Fund. It was also agreed that more than one third of the total package (35bn) was to be destined to recapitalization measures in support of the banking system.

After some hesitation, the Irish parliament did ratify the bailout agreement, but the government fell, new elections were set for 25 February 2011 and resulted in the victory of the opposition who had promised to renegotiate the agreement.

The Irish bailout (as that of Greece) did not have an immediate impact on risk premia and interest rates did not fall (nor for other countries). If anything, the Irish crisis had two major consequences. First it discredited completely the results of the banks' stress tests, as in July 2010 only six small banks had not passed the test and Allied Irish Bank and Bank of Ireland, the two largest Irish banks, both passed the test (Anglo Irish Bank was not included in the tests). Second, it did not allay concerns about the sustainability of Irish debt because the interest charged (close to 6%) on the EFSF loans is much higher than the growth rate Ireland could hope to achieve.

The brief review of the chronology of the crisis shows that Greece was just a trigger and the euro zone crisis is in fact a complex tangle of sovereign debt and banking crises.

The Irish experience has shown that even a government with a strong fiscal position (budget surplus during boom and low initial debt level) can become insolvent in the attempt to save insolvent banks. The sequence of events in Ireland is archetypal: a property bubble ending with a bust leaves a massive housing overhang. This leads to huge losses in banks which had fuelled the bubble with excessive lending. As often happens, the local regulators pretend that there is no problem; but as the losses mount investors pull the plug and the risk of collapse of the entire system increases. This is what happened during the late summer of 2010: as banks were shut out of the interbank market and depositors started to withdraw their funds, the Irish government decided to stand behind the banks and put the entire nation at risk, transforming a banking crisis into the second sovereign debt crisis in the euro zone. A third case of crisis has emerged in early 2011. Portugal has not experienced a bubble as Ireland, neither its fiscal stance is as bad as the Greek one, but the overall financial position of the country is extremely weak. Both private and public sectors have been accumulating excessive levels of foreign debt, which international investors are not willing to finance at sustainable rates and hence increasing dramatically the probability of another bail-out.

The Sequence of the European Council and Euro Group Statements in Response to the Crisis

- **16 February 2010:** The Council focuses on the situation regarding government deficit and debt in Greece, adopting:

(continued)

  - an opinion on an update by Greece of its stability programme, which sets out plans for reducing its government deficit below 3% of gross domestic product by 2012;
  - a decision giving notice to Greece to correct its excessive deficit by 2012, setting out budgetary consolidation measures according to a specific timetable, including deadlines for reporting on measures taken;
  - a recommendation to Greece to bring its economic policies into line with the EU's broad economic policy guidelines.

  http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ecofin/112912.pdf
- **2 May:** Eurozone finance ministers agreed upon a rescue package for Greece amounting to €110 billion: €80 billion in bilateral loans over three years and h30 billion coming from the International Monetary Fund.
- **9/10 May:** The Council and the member states decide on a comprehensive package of measures to preserve financial stability in Europe, including a European Financial Stabilization Mechanism, with a total volume of up to €500 billion from euro area countries and European institutions and the IMF commitment to provide funding up to EUR 250 billion.
  http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ecofin/114324.pdf
- **29 October 29:** The European Council endorses the report of the Task Force on economic governance. The report also sets out the guiding principles for a robust framework for crisis management and stronger institutions; this includes the involvement of the private sector in the crisis mechanism.

(continued)

http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/117496.pdf
- **28/29 October:** The European Council agrees on the need to set up a permanent crisis mechanism to safeguard the financial stability of the euro area as a whole. Eurogroup Ministers agree that the European Stability Mechanism (ESM) will be based on the European Financial Stability Facility, capable of providing financial assistance packages to euro area Member States under strict conditionality functioning according to the rules of the current EFSF. Two further elements are key here:
  - First, support will be available only on the basis of 'a rigorous debt sustainability analysis conducted by the European Commission and the IMF'.
  - Second, 'an ESM loan will enjoy preferred creditor status'.

  http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ecofin/118050.pdf
- **28 November: (Euro-group statement on the Irish rescue package)** Ministers unanimously agreed to grant financial assistance in response to the Irish authorities' request on 22 November 2010. Ministers concur with the Commission and the ECB that providing a loan to Ireland is warranted to safeguard financial stability in the euro area and the EU as a whole. The total size of the package is h85 billion, one-third of it coming from the IMF.
  http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ecofin/118051.pdf
- **16 December:** The Council agreed on the text of a limited amendment to the Treaty on the establishment of a future permanent mechanism to safeguard the financial stability of the euro area. This

(continued)

amendment should enter into force on 1 January 2013. Heads of state reiterated their commitment to reach agreement on the legislative proposals on economic governance by the end of June 2011, with the aim of strengthening the economic pillar of the EMU.

- **24/25 March:** The Council endorses the features of the EMS decided by the euro area Heads of State or Government and takes necessary steps to ensure that the effective lending capacity of the EMS is of EUR440bn. http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/120296.pdf

The Root Causes of the Crisis: Leverage and Bubbles

The euro zone crisis is certainly not the result of a single cause but the outcome of a combination of several factors and dynamics of macroeconomic, regulatory and institutional nature. These include irresponsible behaviour by several euro zone governments, the steady deterioration in peripheral EMU Member States of macroeconomic fundamentals to levels inconsistent with long-term EMU participation, failures in financial market regulation at global level, shift in markets' expectations induced by the financial crisis of 2007–08 and finally, also, defects in the institutional organization of the European Monetary Union.

All these factors are likely to have played a role in originating the crisis, but even together they are still insufficient to account for its systemic nature. This feature can only emerge from the vulnerability of the highly integrated European financial system. Had the Greek and Irish crises occurred when euro zone banks were strong and/or not very interconnected, the euro zone crisis would not have happened. But the European financial system was (and still is) fragile because of the high level of leverage accumulated over the credit boom.[3]

Excessive leverage is an essential ingredient in any major financial crisis and this case is no exception. In financial markets, leverage is defined as the ratio of debt to equity financing; when this ratio increases in general the capacity of a firm to absorb losses declines and hence its fragility is boosted. In macroeconomic terms, leverage is better defined as the ratio of debt to GDP and the concept can be applied to all the sectors of the economy. Leverage defined this way increases when credit expands without a consistent adjustment in GDP. Since regular cash flows are proportional to GDP, this implies that many agents have issued promises to pay a certain nominal amount but do not necessarily have the 'expected' regular cash flow to honour these promises (see Minsky (2008) for the classical description of leverage schemes leading systems towards instability). It is not possible to establish an absolute benchmark for leverage, as different financial systems can support quite different ratios of credit to GDP. However, rapid and persistent increases in this ratio constitute alarm signals which have been identified as reliable predictors of financial crisis. These signals were clearly blinking before 2007, but they were ignored. Table 1 shows that over the last decade euro zone private debt relative to GDP increased by about 100 percentage points, more than it did in the USA. In addition, and unlike the USA, the increase took place in the financial system, whose fragility became apparent first in 2008 and then again in May 2010.

The question is why and how this could actually have happened.

[3]We leave aside the question of why the build-up of the credit boom was ignored. Inflation targeting by central banks was probably one key reason. According to Borio and Lowe (2002), a low-inflation environment increases the likelihood that excess demand pressures show up in the form of credit growth and asset prices bubble rather than in goods price inflation. If this is the case, inflation-targeting central banks with a 'myopic behaviour' could contribute to financial instability (de Grauwe 2009; de Grauwe and Gros 2009).

Euro Zone Crisis 2010, Table 1 Leverage: euro zone versus USA (*Source:* Federal Reserve, Flow of Funds Z1 (outstanding debt), Eurostat and authors' calculations)

Euro area	Non-financial corporations	Financial corporations	General government	Households
1999	67	66	74	49
2007	93	111	69	62
2010	102	127	87	65
US				
1999	63	76	51	67
2007	74	113	51	96
2010	75	101	76	92

Note: For the euro area debt is computed as sum of loans and securities other than shares, excluding financial derivatives (only loans in the case of HH). This definition broadly corresponds to the definition of the outstanding debt used in the US flow of funds

Euro Zone Crisis 2010, Table 2 Leverage for euro zone selected countries and sector break-down (*Source*: Eurostat and authors' calculations)

	Financial corporations		Non-financial sector		Households and non-financial corporations	
Debt-to-GDP	2000	2007	2000	2007	2000	2007
Greece	132	162	175	219	55	105
Ireland*	450	1142	181	294	151	210
Spain	164	310	187	255	122	214
Germany	273	293	200	196	139	130

Note: Debt is computed as sum of loans and securities other than shares, excluding financial derivatives, only loans in the case of households and including also deposits in the case of financial corporations
Non financial sector includes households, non financial corporations and government
*Data for 2000 are not available, those shown refer to 2001

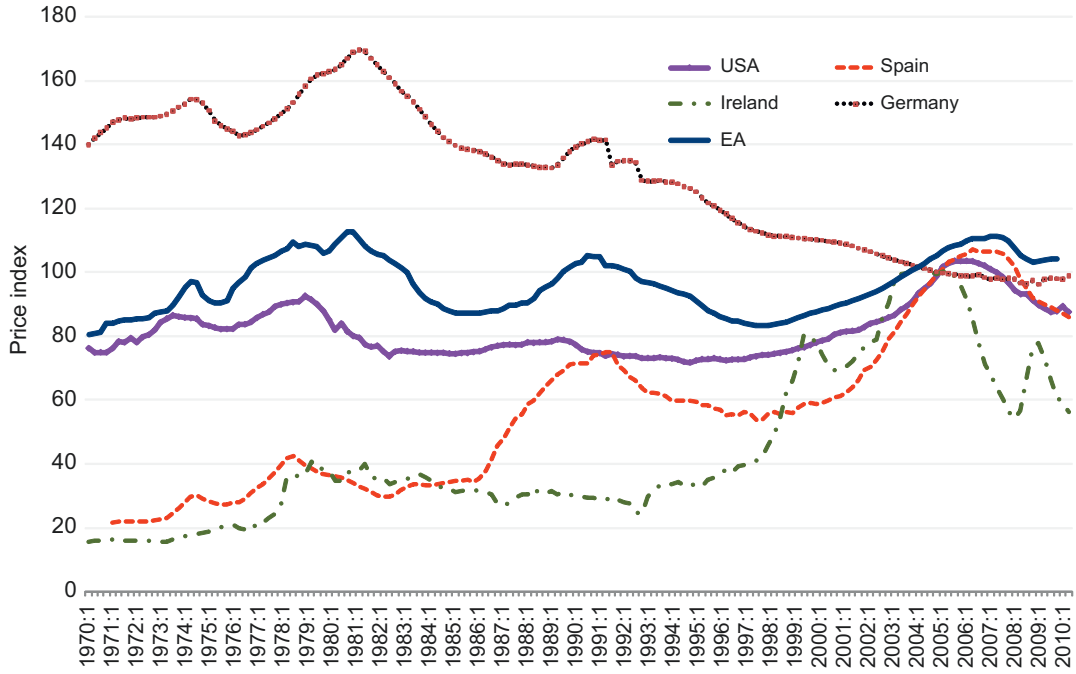
Excess leverage in the banking sector was probably encouraged by scant financial regulation, but it would be too easy to blame car accidents for the absence of speed limits (despite speed limits helping to reduce accidents) or police control. The main driver of growing leverage was of an economic nature and tightly linked to large capital flows flying from core euro zone countries into the periphery after the creation of the euro. The peripheral euro zone economies (Greece, Ireland and Spain) in their catching-up phase appeared to core European Member States with large savings and little domestic investment prospects as a great investment opportunity: they seemed to offer the opportunities of emerging economies, but without the exchange rate risk.

The capital inflows generated their own fundamentals: high growth rates driven by strong demand for consumption and construction investment, supported by easy credit fed from abroad. In all this the financial system, banks in particular, played a crucial role. They made the capital flows possible and magnified the availability of credit through leverage by generating a tight network of intra-sector exposures.

Table 2 shows the level of leverage and the break down by sector in the euro zone countries embedding the most extreme conditions. Data suggest that while leverage barely changed in Germany over the prior decade, in the peripheral euro zone countries, and in particular in Spain and Ireland, the increase was dramatic.

However, it turned out that growth was unsustainable because it was driven by a bubble, and when the bubble burst, banks, not only in the periphery but also in core countries, who were at the origin of the credit flows, found themselves weak (because of high leverage) and very exposed to large potential losses.

The magnitude of the losses was, and still is, potentially very large because some euro zone member countries (notably Ireland and Spain) experienced a real estate price bubble of the magnitude of the USA. Figure 1 provides evidence of



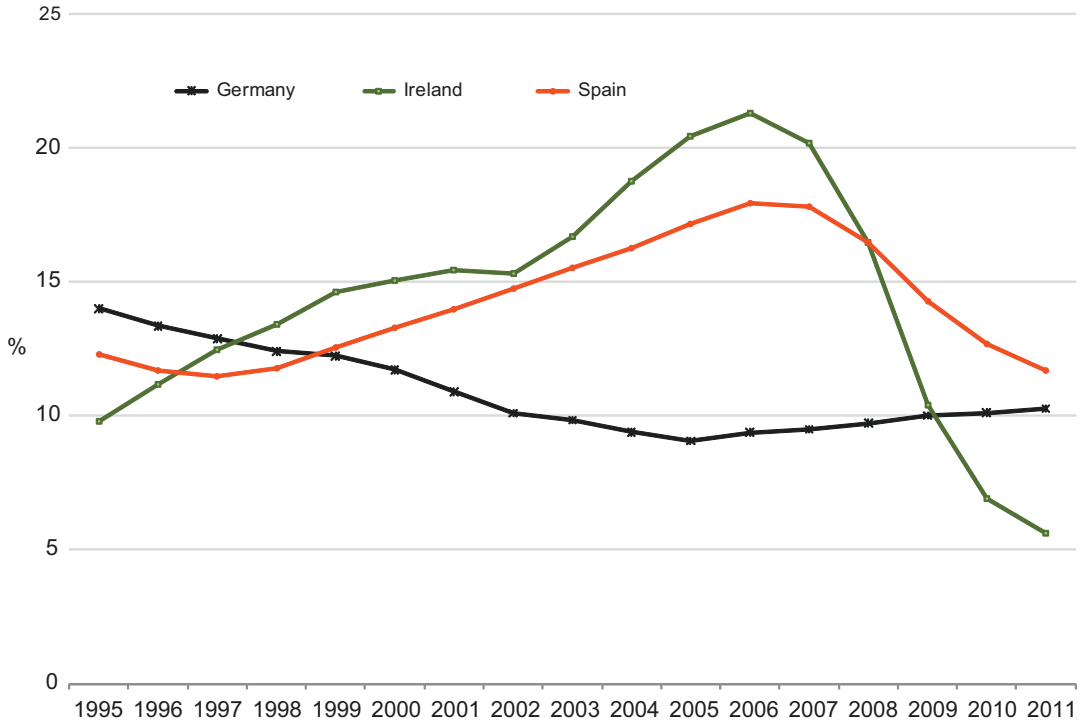
Euro Zone Crisis 2010, Fig. 1 House prices: price-to-rent ratios. *Source*: OECD, March 2011, and author's computations. *Note*: Euro area index is defined as the weighted average (by GDP) of Belgium, Finland, France, Germany, Greece, Ireland, Italy, Netherlands and Spain

this by showing the house price-to-rent ratios. This ratio, similar to the price/earnings ratio for stocks, should be stable over long periods. From the chart it is apparent that since the mid-1990s house prices have increased by almost exactly the same relative amount, reaching an unprecedented level on both sides of the Atlantic. The main difference between the USA and the euro area is that since 2006–07 house prices have declined more in the USA than in the euro zone.

As shown in the Fig. 1, the euro area average hides important differences between countries: Between 1995 and 2006, while house prices have been declining or stable in Germany, they increased by over 80% and more than 140% (more than in the USA) in Spain and Ireland respectively. Furthermore, as shown in Fig. 2, in these two countries the average investment in construction relative to GDP reached 18% and 21% of GDP respectively against an EU average of about 11% (see Fig. 2). This seems to suggest that those countries are destined to suffer for years the consequences of housing and debt overhangs, and dealing with the legacy of national real estate bubbles and busts will remain a challenge for monetary union for some time to come.

This argument is of course related to the so-called 'Walters critique', which holds that a monetary union can amplify shocks because in a country subject to an inflationary pressures the real interest rate will be lower than in the rest of the union. This will fuel domestic demand, which in turn drives inflation even higher, thus lowering real rates even further. This feedback loop is self amplifying and could even be explosive.

However, it seems that in reality the importance of lower real interest rates, defined as nominal interest rates deflated by consumer price inflation, has been overrated. In the case of Spain, consumer price inflation was about 1.6% higher than in Germany over the first 8 years of the euro, but mortgage interest rates were actually over 1% point lower than Germany because they were indexed on short term rates and, even more importantly, house price inflation was 10% points higher than in Germany. This suggests that difference in the characteristics of national financial markets (e.g. the availability of mortgages



Euro Zone Crisis 2010, Fig. 2 Investment in construction % of GDP. *Source*: European Commission Services (Ameco database, Gross fixed capital formation at current prices: construction)

indexed on short-term rates, different loan-to-value ratios etc.) meant that the easing of financial conditions after the creation of the euro had quite differentiated impacts on different member countries (Gros 2009; Baldwin et al. 2010; Calza et al. 2009) with the housing markets playing a key transmission mechanism in Spain and Ireland.

See Also

- Banking Crises
- Euro
- European Central Bank
- European Monetary Union
- Sovereign Debt

Bibliography

Agnello, L., and L. Schuknecht. 2009. Booms and busts in housing markets: Determinants and implications. *ECB Working Paper No. 1071*, European Central Bank, Frankfurt.

Alcidi, C., and D. Gros. 2009. Why Europe will suffer more. *Intereconomics,* July/August.

Baldwin, R., D. Gros, and L. Laeven (eds.). 2010. Completing the Euro Zone rescue: What more needs to be done? e-book, *VoxEU*, June 17.

Blundell-Wignall, A., and P. Slovik. 2010. The EU stress test and sovereign debt exposures. *OECD Working Paper on Finance, Insurance and Private Pensions 4*, OECD Financial Affairs Division.

Borio, C., and P. Lowe. 2002. Asset prices, financial and monetary stability: Exploring the nexus. *Working Paper 114*, Bank for International Settlements.

Calza, A., T. Monacelli, and L. Stracca. 2009. Housing finance and monetary policy. *ECB Working Paper 1069*.

De Grauwe, P. 2009. Keynes' savings paradox, Fisher's debt deflation and the banking crisis. *CEPS Working Document No. 319*.

De Grauwe, P., and D. Gros. 2009. A new two-pillar strategy for the ECB. *CEPS Policy Brief No. 191*.

Eichengreen, B. 2010a. The euro: Love it or leave it? *VoxEU*, May 4. http://voxeu.org/article/eurozone-breakup-would-trigger-mother-all-financial-crises

Eichengreen, B. 2010b. Ireland's rescue package: Disaster for Ireland, bad omen for the Eurozone. *VoxEU*, December 3. http://www.voxeu.org/index.php?q = node/5887

Gros, D. 2009. Comments on Charles Wyplosz, 'Ten years of EMU: Successes and puzzles'. In *Spain and the*

euro: The first 10 years, ed. J.F. Jimeno. Madrid: Banco de España.
Gros, D. 2010. Adjustment difficulties in the Gipsy Club. *CEPS Working Document No. 326*, Brussels.
Minsky, H. 2008. *Stabilizing an unstable economy*. London: McGraw-Hill.
Veron, N. 2010. The EU has not yet solved its banking problem. *VoxEU*, October 26. http://www.voxeu.org/index.php?q = node/5714

Eurodollar Market

W. P. Hogan

The Eurodollar Market is the term still commonly used to describe international financial activities by intermediaries whereby deposits are accepted and loans made in currencies other than the currency of the country in which the participating intermediary is located. The geographical focus was originally in Europe, especially London, but activities are now located in centres around the world.

The origins of the Eurodollar Market lay in exchange control restrictions on the use of national currencies for capital transactions, including the financing of trade between third parties. Banks and other financial intermediaries having longstanding involvement in the financing of such international trade, sought other means of maintaining their position in this sphere. Funds were canvassed in the one capital market where exchange controls did not apply, namely the United States. Thus dollars were borrowed and used to finance exports and imports between countries other than the country in which the financing intermediary was located. Given this circumstance it was inevitable that the US dollar became the currency of denomination for these transactions. Thus the term Eurodollar Market bears witness to the origins of this international financing arrangement. Nevertheless it remains a fair description as more than 70 per cent of all transactions continue to be denominated in that currency.

London remains the largest centre for Eurodollar activity, followed by New York, Frankfurt and Zurich. Transactions may be determined in these centres. The formal completion of transactions may, however, be registered in 'off-shore' places such as the Cayman Islands and the Bahamas in the Caribbean. By completing transactions in various centres external to the United States, such as those in the Caribbean, US banks avoided restrictions on their portfolios such as the locking up of assets in non-income earning reserve requirements. The US authorities bowed to the realities of this situation when in 1981, they established International Banking Facilities in the United States. Extra-territorial recognition to these activities was given while keeping them on-shore.

With the origins of eurodollar activity in a period of rigorous exchange controls, it might have been expected that the market would stagnate, and possibly wither, when exchange controls were whittled down or abandoned altogether as was the case by the late 1970s in many industrial economies. This was not the case. However, distinctions between the Eurodollar Market and transactions in the national currency of the country in which intermediaries were located, were blurred.

Size and Structure

Estimates of the size of the Eurodollar Market or more correctly the Eurocurrency Market are compiled by the Bank for International Settlement, Morgan Guaranty Trust Company of New York and the International Monetary Fund. The IMF series was not published until 1984 so that most attention has been given to the other two. The series of estimates provided by the Bank for International Settlements (BIS) has been the basis for most analysis. The differences between the BIS series and those provided by Morgan Guaranty are explained largely by timing and coverage.

The BIS series is compiled on gross and net bases, the difference being depositing and lending between participating intermediaries. As is evident from the estimates in Table 1, the growth of lending was spectacular for many years during the 1970s. Only in recent times, with the onset of the debt crisis in 1981/82, has lending slowed.

Eurodollar Market, Table 1 Eurofinance lending, 1972 to 1983 (U.S. $ billions)

Amounts outstanding at the end of year				
	Gross		Net	
Year	Value	%increase	Value	%increase
1972	203.7	–	120.6	–
1973	291.4	43.1	172.1	42.7
1974	363.6	24.8	214.6	24.7
1975	442.4	21.7	254.6	18.6
1976	548.0	23.9	324.6	27.5
1977	671.3	22.5	435.0	34.0
1978	856.4*	28.9	530.0*	21.8
1979	1120.3	30.8	665.0	25.5
1980	1335.4	19.2	810.0	21.8
1981	1550.2*	14.1	945.0*	16.0
1982	1694.5	9.3	1020.0	7.9
1983:I	1757.0	0.4	1085.0	6.4
1983:II	2097.9	–	1240.0	–
1984	2153.9	2.7	1280.0	3.2

Source: Bank for International Settlements, various annual reports and reviews.

Note: *Change in series due to alteration in coverage of countries and transactions. Where the series breaks the estimated rate of growth is based on the old series for that year while the new series is the base for estimating growth in the subsequent year. This is illustrated in 1983 where both series are shown.

The difference between the gross and net series for loans outstanding reflects mainly transactions between intermediaries in the Eurodollar Market, often referred to as 'interbank transactions'. The significance of these transactions lies in the provision of liquidity within the market. Intermediaries, by borrowing and lending amongst themselves, could meet liquidity needs arising from the mismatching of maturities between liabilities and assets. Loss of confidence in the asset quality of intermediaries brought illiquidity from 1982 onwards. Many intermediaries, were not prepared to place funds with other intermediaries, a feature evident in the different rates of growth of gross and net lending.

Estimates of the maturity structure of liabilities and assets are provided by the Bank of England for activities in London. About 40 per cent of liabilities have had a maturity of less than a month with over 20 per cent less than eight days. Assets have a longer maturity, about 32 per cent with maturities of less than a month and more than 22 per cent over one year. Maturity mismatching increased during the 1970s and 1980s.

Vital to any understanding of the impact of this market is the short maturities of these deposit liabilities. The financial intermediaries are constantly seeking new deposits or the rolling-over of existing deposits. Given the size of the Eurodollar Market and the frequency with which new funding is required, foreign exchange transactions have come to be dominated by these capital transactions. A modest estimate would be a turnover of US $150 billions per day and, in all likelihood, closer to double that estimate. The scale of these transactions far outweighs transactions related to trade in goods and services.

Mechanisms

Analyses of the workings of the Eurodollar market remain controversial. Initially the market was treated as an extension of a national monetary system; given the predominance of transactions denominated in US dollars, this was viewed as an adjunct of American banking. This general approach for explaining the growth of the Eurodollar Market was matched by the belief that the function of this market was to 'recycle petrodollars'. What that expression meant was simply the functioning of the Eurodollar Market to take up the balance of payments surpluses of the oil-exporting countries after 1973, and then

again in 1980, to fund the deficits of mainly oil-importing countries.

These interpretations do not stand inspection. The Eurodollar Market is not bound by the actions of any one national supervisory authority. Equally, it is not supported by the activities of any central bank. The participating intermediaries, even though they may be called banks, do not have recourse to lender of last resort facilities. They cannot, as yet, write cheques on themselves; rather they must write cheques on accounts in banks within a national banking system. They accept deposits with a specified term to maturity, often very short, and lend for specified periods with provisions for adjusting interest rates. Lending rates are most frequently quoted as some margin over LIBOR – the London Inter-Bank Offered Rate. The participating intermediaries bid for deposits from business, governments, banks, monetary authorities and wealthy individuals all being resident within national monetary systems, and from other participating intermediaries. They lend amongst themselves and to final users of funds, predominantly governments, banks and business in various countries, most often those with trade and payments deficits.

The Eurodollar Market is not an extension of national banking systems. It is a market in debt, not money. The functions performed by the participating intermediaries are central to an understanding of the impact of that market for not only the substantial debts incurred by many countries but also the balance of payments adjustment and exchange rate relativities.

Issues

The stability of the Euromarket mechanism rested upon the capacity of borrowers to meet their obligations. But in providing a means whereby deficit countries could maintain those deficits and not adjust to worsening balance of payments, the participating intermediaries accumulated an increasing proportion of their assets in the obligations of a relatively few countries. Portfolio risk could not be spread.

Impetus to expansion in the Euromarket was maintained during the late 1970s and early 1980s by the narrowing of margins over the cost of funds to participating intermediaries and slender capital/assets ratios. Although the dangers of such developments were recognised quite early in the 1970s, the Bank for International Settlements was unable to make effective its efforts to get coordination amongst national bank supervisory authorities about activities being pursued in the Euromarket. Efforts by the Bank of England and the Swiss authorities, while valuable within their national boundaries, did not gain that wider recognition which, with hindsight, was all too obviously needed.

An explanation for this failure to gain international coordination is the lack of recognition of problems likely to arise with the system. Attention was focused on individual failure either of a participating intermediary or a borrowing country. Only in the 1980s did the systemic problem become clear. By then modest arrangements for coordination proved inadequate, most obviously with the collapse of Banco Ambrosiano in Italy with repercussions for affiliates in Luxembourg and Switzerland.

The rapid escalation of debt problems for chronic borrowing countries in Eastern Europe, Latin America and Africa brought Euromarket activities to a virtual halt by late 1982. Debt renegotiation strained the capital structure of many participating intermediaries and their parent banks. Most were forced to improve capital ratios and hence liquidity, a not surprising development in view of debt rescheduling stretching the maturity structure of assets.

Direct lending by the participating intermediaries meant that the quality of the borrower was not subject to market tests. That lending activity was opaque, not transparent. By maintaining the flow of funds to chronic deficit countries, adjustment problems for those countries were deferred, but the strains were transferred in part to the participating intermediaries. The inherent weakness of this financing system was revealed when rising interest rates, shifts in exchange rate relativities and weak commodity markets in the early 1980s found the chronic debtor countries unable

to service their borrowings and, in many instances, meet repayments.

One repercussion of the harsh strains on participating intermediaries has been to restore activity in international bond financing. That financing, being directly subject to market tests, is confined to countries not facing chronic debt problems. Moreover, governments and companies from those countries are superior credit risks quite often to many participating intermediaries now bearing the penalties of lending to chronic debtors. Those same intermediaries have been willing to foster new techniques of financing, such as note issuance facilities and revolving credits, to maintain participation in international capital transactions. These new techniques offer possibilities for future strains no less than what emerged through direct lending.

See Also

► International Monetary Policy

European Banking Union

Corrado Macchiarelli

Abstract

The sovereign crisis that has characterised the eurozone since 2010 has highlighted the potentially vicious circle between banks and sovereigns, adding an extra dimension to the 2007/08 financial crisis. This is why the EU heads of state and government committed to a European banking union in June 2012; a vision that was further developed in the European Commission's blueprint. The aim of the banking union is to ensure that the financial institutions of the – for now – 19 member states will be subject to a single supervision, a single resolution and a common deposit insurance system. This article explains the background to these initiatives and weighs the progress towards their completion.

Keywords

Banking crisis; Bank resolution; Bank supervision; Economic and monetary union; European Union; European deposit guarantee; Financial crisis; Financial integration

JEL Classifications

G01; G18; G28; O52; E61

European Banking Union

European banking union (henceforth banking union or BU) introduces for the first time an integrated approach in supervision and resolution of European banks, representing an important step towards enhancing economic and monetary union. The aim of banking union is to deliver absolute consistency of implementation of new regulatory rules across the euro area (at the time of writing 129 banking groups, representing more than 80% of the euro area banking sector's assets; ECB (2015)), ensuring that the financial institutions of all member states will be subject to a single supervision, a single resolution and a common deposit insurance system. The need for a banking union emerged in response to the 2007/08 global financial crisis and the ensuing sovereign debt crisis in the eurozone. In particular, the sequence of events highlighted the costs of the vicious link between public and private sector debt, and how these can easily overflow national borders and cause systemic risk and failures.

The banking union proposal dates back to June 2012; it covers a preventive stage (regulation and supervision), and a crisis management stage (resolution and safety nets) (European Commission 2012; IMF 2013b).

The first two components of BU, a single European supervisor (the Single Supervisory Mechanism; SSM) and a single resolution authority (the

Single Resolution Mechanism; SRM) have been agreed. The third element of BU, however – a European deposit insurance scheme covering eligible individual deposits in all participating countries – has been stalling, largely because of political opposition from some creditor member states (Germany in particular).

The SSM has been in place since 4 November 2014; this is the date from which the European Central Bank assumed responsibility for bank supervision. The SSM is a key ingredient of the BU, but it is not the only one. In particular, a European approach to the resolution of banks – with the SRM centred on the idea of a Single Resolution Board (SRB) and a Single Resolution Fund (SRF) – needed to follow. The EU adopted a Bank Recovery and Resolution Directive (BRRD) together with an agreement on the SRM from December 2013. This agreement – following on from the SSM already agreed – was significant because it meant that two of the three components of BU have been operational since 2015. Nonetheless, both elements of BU have been somewhat watered down from their original conception. The SSM will *de jure* not be supervising the whole European banking system, with national authorities continuing to supervise smaller financial institutions. Furthermore, unlike the SSM, the SRM will be 'single in name only' (Posen and Véron 2014) as the framework that sets up the resolution mechanism foresees a significant degree of continuing autonomy for national authorities (see the section below on 'Progress Towards Achieving a Banking Union') – particularly on the issue of funding – at least for the next eight years. Progress has been very uneven on the third element of BU as well, with a common approach to deposit insurance having been sidelined during the first stages of the negotiations. Despite a first legislative proposal for a euro-area wide protection for bank deposits that came as late as 24 November 2015, negotiations are currently stalling. The lack of this third element is critical because it means there will ultimately be no European backstop for depositors in the event of a new banking crisis.

If implemented properly, the original vision for BU may be the most far-reaching reform since the inception of the euro (Constâncio 2013). The fact that the BU vision was further developed in the European Commission's blueprint for economic and monetary union (Juncker et al. 2015) reveals the Eurozone's willingness to continue to deepen integration and to put in place a framework making member states' participation in the eurozone 'sustainable' (see also Pisani-Ferry 2012). However, with an established supervisory authority, a resolution mechanism on the way and a delayed agreement on a common deposit insurance scheme, it remains to be asked whether the European banking project can be credible without a fully fledged fiscal backstop.

Background to Financial Supervision in Europe

Financial market regulation under the Basel Accords, as well as the system of EU financial supervision before 2010, were generally characterised by the lack of mutual recognition. The existing Lamfalussy Process envisaged a largely delegated legislation and enforcement system with an explicit legislation in co-decision procedures (see also ECB 2010). The implementation and transposition of detailed rules on supervision and resolution were delegated to three Committees – the so-called 3 Level Lamfalussy (3 L3) Committees: the CESR (Committee of European Securities Regulators), the CEBS (Committee of European Banking Supervisors) and the CEIOPS (Committee of European Insurance and Occupational Pensions Supervisors). Day-to-day supervision was left to national supervisory agencies, with a strict separation of supervision from central banking, both geographically and functionally (see also Masciandaro et al. 2013).

After 2010, such an approach to financial supervision and regulation changed, under the pressure of the systemic nature of the crisis and the de Larosière report. On the legal side, there was a significant tightening of the regulation of banks, with Basel III raising minimum capital ratios and redefining riskiness of assets. Furthermore, the de Larosière Report (de Larosière Group 2009) established a European Systemic

Risk Board (ESRB), chaired by the ECB's President and Vice-President, with the aim of providing macroeconomic supervision of the financial system as a whole. (For a discussion of the governance of the ESRB see Gerba and Macchiarelli (2015).) The ESRB was created at the end of 2010 as a part of a new two-pillar system of financial supervision, the European System of Financial Supervision (ESFS). The report also gave recognition to three European Supervisory Authorities (ESAs) to cover micro-prudential supervision, representing the ESFS second pillar. These three EU-level bodies, being effective as of 1 January 2011, were not created *ex novo*, but they upgraded the existing 3 L3 Committees. In particular,

- the CEBS was upgraded into the European Banking Authority (EBA);
- the CEIOPS was upgraded into the European Insurance and Occupational Pensions Authority (EIOPA); and finally
- the CESR was upgraded into the European Securities and Markets Authority (ESMA).

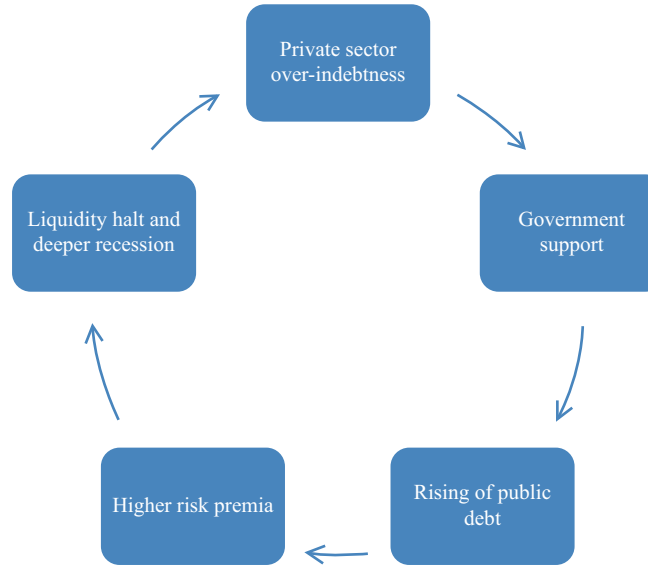
This change in governance structure marked not only the beginning of a greater (than in the rest of the world) involvement of the central bank in Europe, but also the start of a two-pillar strategy ensuring – by means of institutional separation and coordination with national supervisors – a system of checks and balances between macro- and micro-prudential supervision (see also Goodhart and Schoenmaker 1995; Masciandaro et al. 2013; Eijffinger 2013; Goodhart 2014).

The first agreement on banking union came in September 2012. The European Parliament's final 'go-ahead' for the ECB to be fully entrusted with responsibility for the supervision of banks in the framework of the SSM came after extensive negotiations between various stakeholders. This happened one year after the first agreement, on 12 September 2013. The 2012 EU Council agreement appropriately conferred broad investigatory and supervisory powers on the ECB, which – as of November 2014 – is responsible for the effective and consistent functioning of the SSM. National authorities remain responsible for the banks remaining under their direct supervision (the so-called 'less significant financial institutions').

Guidance on the design of an effective supervisory mechanism for Europe was provided in the Basel Core Principles (the so-called 'Core Principles for Effective Banking Supervision'; http://www.bis.org/publ/bcbs30a.htm). According to these principles, a number of preconditions and prerequisites were to be met at the euro area level, including (1) the implementation of coherent and sustainable macroeconomic policies; (2) a clear framework for financial stability policy; (3) an effective crisis management and resolution framework to deal with bank failures and minimise disruptions; (4) an adequate safety net to deal with confidence crisis while minimising distortions; (5) a well-developed public infrastructure; and (6) effective market discipline. On the other hand, as underlined by IMF (2013b), prerequisites to establish a sound basis for the SSM included: (1) its operational independence; (2) clear objectives and mandates; (3) legal protection of supervisors; (4) transparent processes, sound governance and adequate resources; and (5) accountability (see also IMF 2013b; Gerba and Macchiarelli 2015). As we shall discuss in the following sections, after the comprehensive assessment performed by the SSM at the end of 2013, with extensive granular balance-sheet facts being provided and a higher degree of transparency and availability of information to the public, the SSM seems to meet these criteria.

The European 'Doom Loop'

The crisis highlighted the importance of having in place a framework for dealing efficiently and in a timely manner with the resolution of cross-border financial entities (Obstfeld 2013), avoiding the long-term implications on fiscal sustainability of having national governments and banks dangerously tied together (see, *inter alia*, Reinhardt and Rogoff 2013; Gennaioli et al. 2014). These ties essentially intensified during the eurozone's crisis for two reasons. First, banks engaged in *carry-trade* by using 'cheap' central bank liquidity to purchase government bonds (Acharya and Steffen



European Banking Union, Fig. 1 A stylised representation of the European 'doom-loop'

2015). (Central bank liquidity came mainly in the form of three-year long-term refinancing operations (LTRO), with the interest rate fixed at the average rate of the main refinancing operations at the time (1% p.a.) and full allotment of the bids (for further technical details see ECB 2011).) Second, there was a rapid rebalancing of banks' international portfolios towards 'home' assets and bonds (Battistini et al. 2014; Valiante 2015). The latter was possibly the result of risk-shifting (Gennaioli et al. 2014; Farhi and Tirole, 2016; Acharya et al. 2015); discrimination (Broner et al. 2016); and financial repression (Chari et al. 2014; for a general discussion see also Reinhart et al. 2011).

Government guarantees to banks at the expense of higher debt and the inability of regulators to stall the crisis, together with a 'faulty' design of the currency union – centred on a single central bank and multiple Treasuries (see De Grauwe 2016) – are known to be amongst the weighty factors at the root of the private-public European 'doom loop'.

In the euro area, in particular, together with the impossibility of monetising debt (an explicit provision contained in the Lisbon Treaty on the Functioning of the EU – the 'no bailout rule' – art. 125), as of 2010 countries had *de facto* to compete 'internally' over capital flows (Valiante 2015). This was the reflection of an institutional setup built on the idea of 'tying one's hands' – i.e. guarding against government failure by simply agreeing on strict fiscal rules (e.g. the Stability and Growth Pact) and letting markets find their equilibria (Fuest and Peichl 2012). (A key reason for the failure of international capital markets to differentiate sufficiently between countries according to the state of their public finances was that the 'no bailout rule' was just not credible (Fuest and Peichl 2012). In other words, before 2010 financial markets simply did not set incentives to limit government debt in the Eurozone, and very small borrowing premia were to be paid over German *safe-haven* rates.) The sovereign debt crisis that followed confronted almost all non-AAA-rated euro area countries (Greece and Ireland first, followed by others by 2012) with a liquidity dry out, as the result of a flight-to-quality of capital – facilitated indeed by the single currency – towards their 'safer' EMU peers. This translated into higher public borrowing costs, a frailer banking system and overall larger bailout charges *ex post*.

Figure 1 proposes a stylised representation of the aforementioned European doom-loop. This representation does not consider contagion or spillover effects from, or to, other countries, being broadly related to recent literature on doom loops in closed economies (see e.g. Acharya et al. 2015). In this representation, whatever the

entry point is (private sector leverage, unsustainability of public finances, lack of structural reforms) there is a self-reinforcing effect relating to the classical problem of (ir)rational runs in which the market can push an economy into a 'bad' equilibrium (see also De Grauwe and Ji 2013). This amplification within the EMU had to do, firstly, with a collapse of confidence in certain markets and institutions at the same time, and the broader fragility of financial systems, because of increased counterparty risk or asymmetry of information (see also IMF 2013a). Secondly, it was linked to the distressed financial sector inducing government bailouts (or private sector *deleveraging*; see Acharya et al. 2015). The latter, in particular, created a vicious interaction between asset prices (via banks' balance sheets) and borrowing constraints (Borio and Zhu 2012; Brunnermeier and Pedersen 2009; De Grauwe and Macchiarelli 2015), where – simplifying – the fire sale of government bonds in some countries (as the result of confidence loss and excessive debt taking) increased sovereign credit risk, in turn weakening the financial sector, with an ensuing liquidity dry out and freezing of lending to the real economy. Overall this eroded bond holdings and the value of government guarantees, requiring further support, and so on.

Why a Banking Union for Europe?

The governments' last-resort guarantees to their own financial institutions were initially granted in an uncoordinated manner within the EU. Government asset support mainly took two forms (see also Gros and Schoenmaker 2014): asset insurance schemes, which maintained the assets in the banks' balance sheet, and asset removal schemes, which transferred the assets to a separate institution (bad bank). Purchases of impaired assets often occurred after earlier government capital injections. In the case of bank debt guarantees, approximately half of those that received capital injections also received government guarantees for their bank debts. Coordinated support happened only later and was led by the European Commission in the context of its State Aid policy, with the aim of preserving an EU integrated financial market. Before the Commission launched a bank recovery proposal, a number of EU countries, including Austria, Belgium, Denmark, France, Germany, Ireland and the UK had already put in place new rules for the resolution of their distressed banks. Such repeated bailouts not only increased sovereign debt, but also imposed a large encumbrance on taxpayers. The state aid measures that were used, in the form of recapitalisation and asset relief measures between October 2008 and December 2012, amounted to €591.9 billion or 4.6% of EU 2012 GDP, with the highest share belonging (in order) to Ireland, the UK and Germany (European Commission State Aid Scoreboard's (2013) figures. Available at http://ec.europa.eu/competition/state_aid/scoreboard/amounts_used_2008-2012.xls). Including approved aids and guarantees, this figure jumps to over 12% of EU GDP for the period 2008–12 only. In the euro area, 37% of capital injections and 63% of the asset relief measures were granted to the three largest financial institutions (see also Gros and Schoenmaker 2014).

Beyond government upkeep, central banks provided unprecedented liquidity support to illiquid (and insolvent) banks as well. Specifically, the European Central Bank during the first stage of the crisis focused its programme – albeit not exclusively – on direct lending to banks (see the preceding section), reflecting the bank-centric structure of the euro area financial systems (see also Gabor 2014; ECB 2014). This was different from the Federal Reserve and the Bank of England, which expanded their respective monetary bases largely by purchasing bonds in the first place.

Looking at the recent history of bailouts, the advantage of a permanent bank supervision and resolution framework, as compared to the *ad hoc* measures that were employed during the crisis, primarily resides in its transparency regarding the list of eligible institutions and the conditions of access to funding. Second, it introduces a limitation to free-riding derived from unlimited recourse to public money, allowing overall a balanced burden-share between private investors and taxpayers, possibly resulting in lower funding

costs *ex post*. At the same time, the BU proposal recognises the systemic nature of risk facilitated by the single currency, and the potential dangers and domino effect that 'systemically important' financial institutions would have, given their cross-border reach, within the E(M)U (see also Obstfeld 2013; Gros and Schoenmaker 2014; Goodhart 2014). Finally, the proposal acknowledges the issue of the moral hazard of national governments both over time – with a tendency to offload the costs of restructuring the domestic banking sector to future governments – and across countries – particularly, relying on the ECB's and European Stability Mechanism's last resort support. The latter two points relate to the literature analysing the combination of limited commitment on the part of the government *ex ante*, and the possibility of bailouts *ex post* (see, among others, Acharya and Yorulmazer 2007; Chari and Kehoe 2016; Farhi and Tirole 2012). In particular, this literature highlights a mechanism by which government bailouts are provided only when a sufficient number of financial institutions are in trouble *ex post*, so that strategic complementarities in financial risk-taking arise: i.e. individual financial institutions may engage in higher financial risk-taking *ex ante* the higher the collective risk-taking, as this increases the likelihood of a government bailout *ex post*. The existence of such complementarities and systemic risk thus provides a rationale for macro- and European measures. (Broner et al. (2016) put forward another rationale for a BU: a BU is thought to reduce discrimination between domestic and foreign investors.)

Legal Underpinning

The legal foundation of BU is contained in a single rule book made up of three main elements.

1. A set of **rules on capital requirements** (Capital Requirements Directive – CRD IV).

These rules entered into force on 1 January 2014, and replaced the original Capital Requirements Directives (2006/48 and 2006/49), transposed the international Basel III agreement into EU regulation and ensured that banks hold a sufficient buffer to withstand potential losses.

2. The proposal for strengthening the **Deposit Guarantee Schemes Directive** (DGSD) (Revision of Directive 94/19/EC).

The aim of the latter was to harmonise and simplify deposit guarantee rules in the EU and improve the functioning of the existing guarantees across the board, with protection of deposits up to €100,000 (from the existing €20,000 limit). According to the directive, all credit institutions will be required to join the DGS instituted at the national level. The Council has reached a political agreement with the European Parliament on the revised directive, with the Parliament formally adopting this revision in April 2014.

3. **Bank Recovery and Resolution Directive** (BRRD) (Directive 2014/59/EU).

This directive gives powers to authorities across the EU to act effectively to prevent bank crises and to ensure orderly restructuring and resolution in the event of bank failure. The aim is to avoid negative effects on financial stability and to reduce recourse to taxpayers' money, avoiding replicating the scenario seen during the first stage of the crisis. The directive followed a Commission proposal in June 2012. The European Parliament and the member states reached an agreement on 11 December 2013. These new rules, which entered into force on 1 January 2015, established that the costs of bank failure will in the first instance be borne by bank shareholders and creditors, according to a clearly defined hierarchy, and thereafter met from dedicated resolution funds held by each member state.

Progress Towards Achieving A Banking Union

Common Bank Supervision

A European single supervisor (SSM) became operational in November 2014 (see section on 'Background to Financial Supervision in

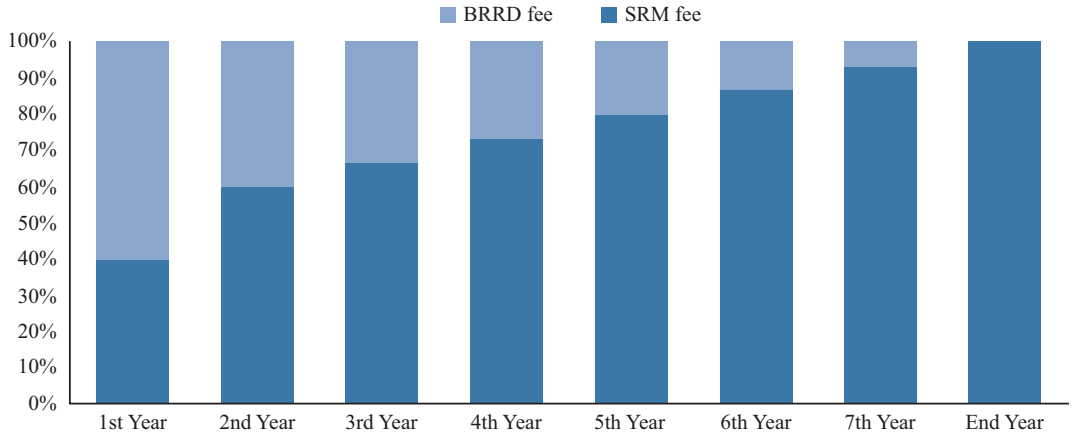
Europe'). Under the SSM, responsibility for bank supervision in the euro area was shifted from national authorities to the European Central Bank. The ECB is in charge of supervising 'systemically important' banks directly (equal to more than 80% of euro-area banking assets, including banks with over €30 billion in assets or 20% of national GDP, or 'if otherwise deemed systemic'). National authorities will continue to supervise smaller financial institutions. (In September 2014, the ECB published the list of significant supervised entities. The latest release (31 May 2016) with change in significance for some banks is published here: https://www.bankingsupervision.europa.eu/ecb/pub/pdf/list_of_supervised_entities_20160331.en.pdf.) The latter arrangement was essentially a political one, championed by some member states – Germany primarily – wanting to keep direct monitoring of 'local' institutions. The federal approach that emerged as a concession to local banks' lobbies highlights how banks' management in some countries cultured a strict affiliation with the political establishment and local electorate (Valiante 2015), largely through 'not-for-profit' credit institutions such as foundations (e.g. the Spanish Cajas and the German Landesbanken). Overall, however, while smaller banks were *de jure* exempted from direct SSM supervision, the €30 billion threshold has *de facto* left the majority of the eurozone banking assets under the SSM's umbrella – including almost all German Landesbanken (Posen and Véron 2014). Furthermore, the ECB will set and monitor supervisory standards and work closely with the national competent authorities for these banks, with the option of expanding its remit and supervising them directly in order to ensure that SSM standards are applied consistently (ECB 2014, 2015).

To conclude, while the design of a common bank supervisor is far from faultless, given the challenge to financial stability that small financial institutions may pose, these challenges in terms of supervision are, for the moment, not large. (It is worth noting that the majority of 'local' banks in the EMU are concentrated in Germany, and, to a lesser extent, Austria and Italy; see Véron's blog entry on Bruegel: 'Europe's Single Supervisory Mechanism: Most small banks are German (and Austrian and Italian)', 22 September 2014.) The current SSM design represents an adequate compromise given the existing trade-off between political feasibility and economic 'first-best' in Europe. In addition, achieving a truly single market in banking services will possibly require more time than a couple of years, with further supervisory initiatives, as well as regulatory and legislative steps, having to be adopted in the future (see also Schoenmaker and Véron 2016).

The Single Resolution Mechanism

The SRM was first proposed by the European Commission in July 2013. This mechanism came to complement the SSM as of 1 January 2015. Countries joining the SSM are to join the SRM too, which means that the SRM applies to banks in the euro area member states under the SSM, plus those EU countries wishing to opt in. The SRM is built on the national resolution authorities established by the BRRD. The SRM aims to ensure that if – despite SSM supervision – a bank faces serious difficulties, its resolution would be managed in a centralised manner, with minimal cost to taxpayers and the real economy; which is one of the focal points of BU.

The SRM consists of a resolution authority (or Single Resolution Board – SRB) and a Single Resolution Fund (SRF). The SRB became operational in January 2015, but it started to work at full capacity one year later, on 1 January 2016, the date when the SRF was also on the schedule. The Finance Ministers of the member states have decided to keep some elements of the functioning of the future SRF in the form of an intergovernmental agreement, which complements the SRM regulation. According to the terms of reference of the agreement, the fund is to be financed by bank levies raised at the national level. As a general rule, banks taking higher risks will pay higher contributions. Contributions, initially consisting of national compartments, which will be progressively mutualised and eventually merged into a single fund administrated by the Board, start with 40% of resources in the first year. National



European Banking Union, Fig. 2 Evolution of phasing-in (−out) of contributions to SRF (from national target levels in accordance to the BRRD) (*Source*: ECB (2015) data)

compartments would cease to exist when the fund reaches the target funding level of 1% of covered deposits in the participating member states or after an eight-year transitional period – i.e. by 2024.

Under the SRM Regulation, the SRB is required to calculate the contribution from each individual bank to the SRF each year. Contributions are determined by applying the method detailed in a Commission delegated act and the specifications provided for in a Council implementing act, adopted respectively on 21 October and 19 December 2014. The establishment of the SRF will thus entail a shift from national to European resolution, which has the implication that each member state's banking sector will progressively contribute more to the European resolution fund with respect to what they will be contributing to the national fund under the BRRD. This is summarised in Fig. 2.

The SRF has an overall target level of €55 billion. While this amount may seem small in principle – given the need to signal to the markets that a reliable backstop exists (see also Macchiarelli 2014; Gros and Schoenmaker 2014) – one should consider that the fund has been given the ability to borrow directly from the market, if decided by the Board (ECB 2015); the terms and conditions of this have not been disclosed yet. Secondly, explicit provisions for bailing-in exist, as detailed by the revised BRRD (SRF website, https://srb.europa.eu/en/content/bail (accessed August 2016)). Bailing-in would apply until at least 8% of banks' total assets had been used. After this threshold, the resolution authority may grant the bank the right to use the resolution fund, up to a maximum of 5% of the bank's total assets. Some have observed how the actual procedures for bailing-in may not only risk cutting credit in already fragile economies, but could also reduce the willingness of lenders to extend new credit, having overall a negative effect on the financial conditions of that country.

Bank contributions to the SRF began in January 2016. However, a plan on bridge financing was put in place in the context of the Five Presidents' Report (Juncker et al. 2015) in order to avoid a situation in which the SRF would run out of monies while bank contributions were being consolidated. The agreement, which was reached by the Council of Ministers in December 2015, introduced public support through the establishing of national credit lines that would provide a loan to the SRF in the case of capital shortfalls before 2024. As well as providing support where needed, the establishment of credit lines is intended to enhance the standing of the SRF. Importantly, a common backstop to the SRF itself should follow before the end of the transitional period, as a last resort measure, in order to ensure the durability of the BU project as a whole, as the Five Presidents' Report also recognises (Juncker et al. 2015). (See also Communication

from the Commission to the European Parliament, the Council, the European Central Bank, the European Economic and Social Committee and the Committee of the Regions, 'Towards the completion of the Banking Union' COM (2015) 587. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri = CELEX:52015DC0587.) This will be difficult to achieve in the short term as it will require a more far-reaching fiscal agreement among the member states (for an extended discussion see the following section).

Banking Union and Fiscal Backstops

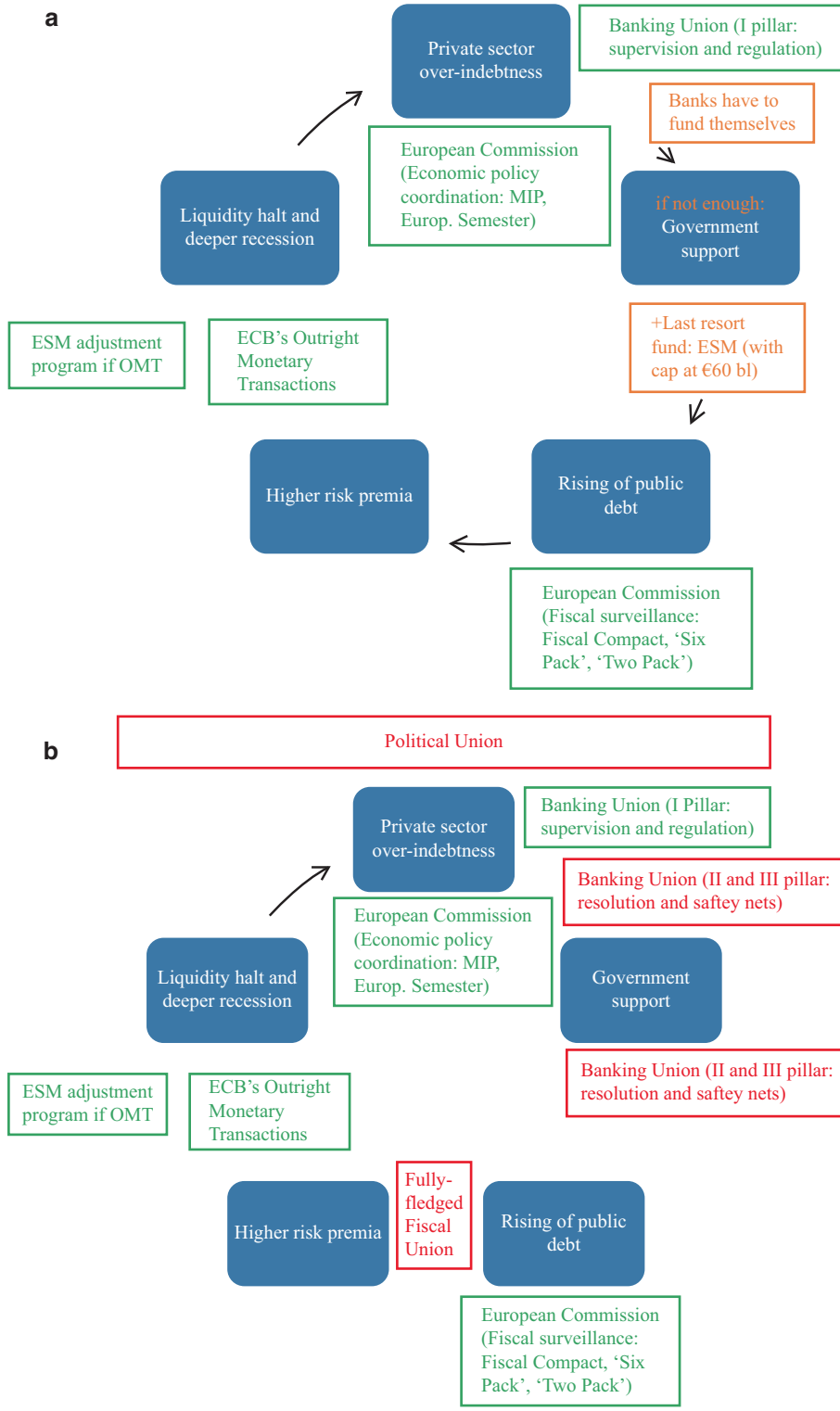
While it is widely recognised that the proposals and compromises reached to deal with deposit insurance and resolution represent an exceptional step forward, many member states underscore that a well-functioning BU will require an unlimited burden-sharing mechanism, where fiscal authorities have to be involved. As highlighted above, the current design of the BU still leaves a role for an intergovernmental agreement, particularly in deciding the role and functioning of the future SRF – as a complement to the SRM regulation – before its final consolidation by 2024. Furthermore, the Commission's deposit insurance mechanism (EDIS) is still on the negotiating table. Hence the stage in the governance framework that is lacking is the fiscal backstop.

The existing national DGSs and resolution funds – before a common backstop is created – may quickly run out of money and need last-resort support from sovereigns. This, however, was the origin of the so-called doom loop, pushing even countries with a sound fiscal record into a wrecking spiral, as the cases of Ireland and Spain show. Should this be the case, the sovereign will then need a backstop itself. In Fig. 3a, b we have used the 'doom loop' representation of Fig. 1 to summarise this discussion. The figure particularly compares (a) the current state of BU with a representation of (b) fully fledged BU in the context of the GEMU.

The current state of BU is an incomplete banking union which could create coordination failures and be costly overall (Posen and Véron 2014). As mentioned above, leaving resolution funding and safety nets predominantly at the national level or, equally, limiting the BU's 'federal' reach and burden-sharing capacity, would mean perpetuating the bank–sovereign doom loop, which is what the BU is intended to break. Alternative arrangements exist, but it remains to be seen whether these are convincing.

Figure 1(a) highlights a role for the European Stability Mechanism (ESM), as a last resort support. The ESM was primarily created with the purpose of providing a fiscal backstop for member countries and (more recently) their banking systems. However, the stability of banking systems can be assured only if investors know that such a backstop is not limited *ex ante*. This is why many commentators have defined the ESM as a 'poor surrogate' for a last-resort lender (De Grauwe 2011). The main reason why the European banking sector needs a common backstop is fundamentally macroeconomic and has to do with the very nature of systemic risk (Gros and Schoenmaker 2014; see also Allen et al. 2011). Once all of the above is in place (SRF plus EDIS), in the great majority of cases no public support will be needed. But in exceptional circumstances, for relatively large shocks, additional resources might be necessary, and clear arrangements on backstops should be made. Thus, the stability of banking systems can be assured only if investors know that such a European backstop exists, and the current ESM capacity can hardly be credible (see also Gros and Schoenmaker 2014).

Secondly, there are agency costs to consider, as the ECB/SSM may itself be trapped in a fiscal dominance game (see also Goodhart and Schoenmaker 1995; European Parliament 2012). The existence of a transition period before the SRM and the EDIS (whose deadlines for full functioning are aligned) makes it possible that, until resources are fully mutualised, the SSM will have an incentive to offload the fiscal cost of any problem to national authorities if it thinks that any given bank is insolvent and needs to be restructured or closed down. The SSM would do this on the basis of its comprehensive assessment of the viability of the bank and any danger it might constitute for financial stability. By contrast,



European Banking Union, Fig. 3 (continued)

national authorities in charge of bank restructuring would have a tendency to minimise their own costs by keeping the bank (even if illiquid) solvent through ECB support. This leaves some grey areas in the crisis management capacity of the BU (ECB 2015), with this type of conflict being prevalent between now and the start of the new system, when mutualisation is low. The endgame would be accelerating the process of consolidation of European deposit insurance and resolution schemes, thus minimising potential costs and avoiding providing the SSM and national authorities with the wrong incentives. (Other inter-agency conflicts and fiscal dominance may arise in the context of keeping two different coffers for European deposit insurance and resolution, respectively (for an extended discussion see Gros and Schoenmaker 2014).)

The nature of fiscal backstops beyond resolution and safety nets will be a crucial issue to define in the coming years.

Safety Nets

Authorities are now equipped with a broad set of tools to ensure that the costs of bank failure will, in the first instance, be allocated to bank shareholders and creditors following a clearly defined hierarchy (bailing-in), and only later involve dedicated resolution funds held at the national level (bailing-out). (Higher coverage will be granted for deposits related to certain transactions (e.g. real estate transactions and payment of insurance benefits). See ECB (2015).) In particular, as far as deposit protection goes:

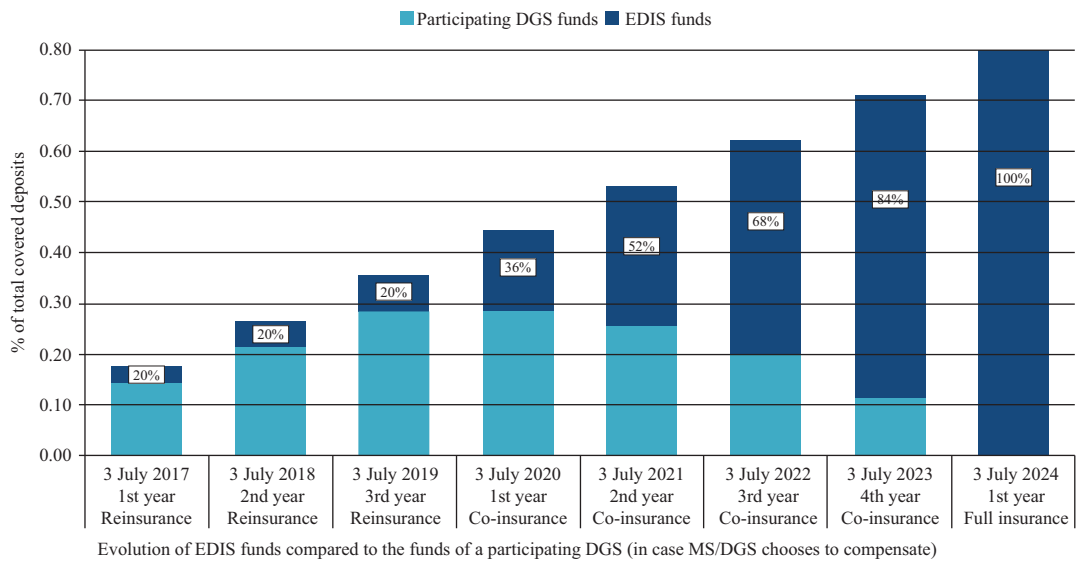
- Citizens' covered deposits up to €100.000, representing about 48.6% (47.3%) of total euro area (EU) deposits, will be exempt from any loss. This number goes up to 70.9% (66%) for the euro area (EU) if the eligible over total deposits ratio is considered (author's computation from Cannas et al. (2014) data).
- Deposits of natural persons and SMEs above €100.000 will benefit from preferential treatment (they will not suffer any losses before other unsecured creditors do).

The Deposit Guarantee Schemes Directive (DGSD), which was transposed by the member states into national law in July 2015, concentrated on harmonising existing national deposit guarantee schemes without any common funding element. While regulators agreed to an increase in the minimum coverage of insured deposits from €20,000 to €100,000 and an increase in the speed of repayment for insured depositors, the most worrying gap is that of the unification of deposit insurance within the banking union.

In November 2015 the Commission put forward a legislative proposal to fill this gap, i.e. a European deposit insurance scheme (EDIS), taking a concrete step towards completing the third leg of BU. This is a very significant proposal, as the absence of a union's deposit guarantee that could credibly back it underscores the dangers of incompleteness. A DGS funded at the European level would, in this case, make a material difference because it would provide an external loss absorption device that would be independent of the fiscal position of that sovereign (Posen and Véron 2014). Yet the EDIS has still to be approved, and it is currently stalling owing to political opposition.

The DGSD stipulates new thresholds for the financing of the national Deposit Guarantee Scheme (DGS), notably by requiring a significant level of *ex ante* funding (0.8% of covered deposits – or, where viable, a target level of 0.5% of covered deposits for highly concentrated banking systems) to be built up by 2024 by each member state. By that date, the Commission's proposal envisages that resources will be mutualised in the EDIS. With the EDIS, the protection of deposits

European Banking Union, Fig. 3 European banking union. (**a**) Current state of European banking union; (**b**) European banking union in the context of the GEMU – in theory (*Note:* The figures include the main reforms of the European economic governance framework already in place (*green*); measures to be adopted during the transition to a BU (*orange*), and measures not yet in place (*red*). They do not consider measures which are temporary in nature such as unconventional monetary policy)



European Banking Union, Fig. 4 Evolution of EDIS and participating DGSs funds in the Commission's proposal (*Source:* European Commission website – European deposit insurance scheme)

would be fully guaranteed at the European level, supported by close cooperation with national DGSs (Fig. 4). Given that national DGSs may remain vulnerable to idiosyncratic shocks, the purpose of EDIS would be to ensure equal protection of deposits in a centralised manner.

In the period elapsing between now and the EDIS, harmonised national deposit guarantee schemes will be necessary, meaning that member states concerned by a particular resolution plan will have to provide bridge financing from national sources (ECB 2015). In particular, if capital shortfalls are identified, the Council clarified on 15 November 2013 the order of the backstops. In the case of insufficient *ex ante* funds, the DGS must collect *ex post* contributions from the banking sector. Exceptional contributions should not exceed 0.5% of covered deposits per year. In the first instance, banks will thus have to raise capital in the market or raise capital from another private source. Should this not be sufficient, public money could be engaged at the national level in line with state aid rules and, if needed, through the provision of public backstops. Here, the DGS may have access to alternative funding arrangements, such as loans from public or private third parties. The DGSD also establishes a voluntary mechanism of mutual borrowing between DGSs from different EU countries (ECB 2015), the viability of which has still to be tested, particularly given the possibly competing interests of debtor and creditor countries.

Should national backstops not be sufficient, instruments at the European level may finally be enabled, including the ESM, consistent with the ESM's agreed procedures. On the latter point, the Eurogroup agreed that the ESM would have the possibility to recapitalise 'systemic and viable' banks directly, with maximum exposure for direct bank recapitalisations capped at €60 billion (equal to 12% of the ESM's maximum lending capacity).

Given the uncertainty about the full viability of the project in the medium to long term, information to markets and depositors should be prepared and coordinated.

Managing 'The Outs'

One issue with the current approach to the European BU is that it minimises the importance of cross-border externalities of bank failures across the EU. Given the skewed design of the BU towards the euro area member states, the problem of funding is likely to be more severe when it

involves guarantees to or resolution of banks which are systemic in both euro area and EU-non-euro area countries. For that reason, some of the 'outs' may make use of their option to opt-in to BU going forward (Gros and Schoenmaker 2014), provided that European resolution and deposit insurance schemes will be available. In this respect, the UK's vote to leave the EU will place both the EU and the UK in uncharted waters, given the large presence of important European banks in London, and in the absence of clear rules on cross-border banking supervision and resolution under BU across EU and non-EU member states.

Final Remarks

A European banking union centred on the idea of single supervision, single resolution and a common deposit insurance system may be the most far-reaching reform to date since the inception of the euro (Constâncio 2013), if successful. Overall, however, the political resistance of creditor countries may restrain the effectiveness of crucial elements of BU, such as resolution and safety nets. A credible banking union would entail moving responsibility for potential financial support from the national to the supranational level, implying transfer of resources and risk, and, henceforth, requiring an explicit agreement on fiscal European support in the longer term. The latter agreement is a necessary step in the broader context of the European governance framework (see ▶ Genuine Economic and Monetary Union), in particular in achieving long-term 'sustainability' (see also Pisani-Ferry 2012; Gros and Schoenmaker 2014; Posen and Véron 2014; Schoenmaker and Véron 2016). For the time being, political resistance mainly focuses on the issue of permanent and unlimited vs. temporary and limited burden-sharing, leading to a 'small steps' approach.

An incomplete banking union can create coordination failures and could be costly overall (Posen and Véron 2014). An incomplete union can be interpreted in two ways. One is that it is a sequence in which much remains to be settled, but with reasonable clarity about the eventual destination. In this interpretation, the principal policy challenges will be how to manage the transition until 2024. The alternative explanation is that political resistance to burden-sharing will mean that only an incomplete banking union can be attained in fact. As mentioned above, leaving resolution funding and safety nets predominantly at the national level – i.e. the current state of BU – or, equally, limiting the BU's 'federal' reach and burden-sharing capacity, would mean perpetuating the bank–sovereign doom loop; which is what the BU is intended to break. The nature of fiscal backstops beyond resolution and safety nets will be a crucial issue to define in the coming years.

See Also

- ▶ European Central Bank
- ▶ European Monetary Union
- ▶ Genuine Economic and Monetary Union
- ▶ Regulatory Responses to the Financial Crisis: An Interim Assessment

Acknowledgments The author wishes to thank, without implicating, Iain Begg for his constructive comments on this entry.

Bibliography

Acharya, V., and S. Steffen. 2015. The greatest carry trade ever? Understanding eurozone bank risks. *Journal of Financial Economics* 115: 215–236.

Acharya, V., and T. Yorulmazer. 2007. Too many to fail – An analysis of time inconsistency in bank closure policies. *Journal of Financial Intermediation* 16(1): 1–31.

Acharya, V., I. Drechsler, and P. Schnabl. 2015. A pyrrhic victory? Bank bailouts and sovereign credit risk. *Journal of Finance* 69(6): 2689–2739.

Allen, F., T. Beck, E. Carletti, P. Lane, D. Schoenmaker, and W. Wagner. 2011. *Cross-border banking in Europe: Implications for financial stability and macroeconomic policies*. London: CEPR.

Battistini, N., M. Pagano, and S. Simonelli. 2014. Systemic risk, sovereign yields and bank exposures in the euro crisis. *Economic Policy* 29(78): 203–251.

Borio, C., and H. Zhu. 2012. Capital regulation, risk-taking and monetary policy: A missing link in the transmission mechanism? *Journal of Financial Stability* 8(4): 236–251.

Broner, F., A. Erce, A. Martin, and J. Ventura. 2016. Sovereign debt markets in turbulent times: Creditor discrimination and crowding-out effects. *Journal of Monetary Economics* 61: 114–142.

Brunnermeier, M., and L. Pedersen. 2009. Market liquidity and funding liquidity. *Review of Financial Studies* 22(6): 2201–2238.

Cannas, G., J. Cariboni, L. Kazemi Veisari, and A. Pagano. 2014. *Updated estimates of EU eligible and covered deposits*. Luxembourg: European Commission Joint Research Centre – Institute for the Protection and Security of the Citizen.

Chari, V.V., and P.J. Kehoe. 2016. Bailouts, time inconsistency, and optimal regulation: A macroeconomic view. *American Economic Review* 106(9): 2458–2493.

Chari, V.V., A. Dovis, and P. Kehoe 2014. *On the optimality of financial repression*. *Mimeo*, Federal Reserve Bank of Minneapolis.

Constâncio, V. 2013. The European Crisis and the Role of the Financial System. Speech at the Bank of Greece conference 'The crisis in the euro area', Athens, 23 May.

De Grauwe, P. 2011. The European Central Bank as a Lender of Last Resort. VoxEU. org, 18 August.

De Grauwe, P. 2016. *Economics of Monetary Union*. 11th ed. Oxford: Oxford University Press.

De Grauwe, P., and Y. Ji. 2013. Self-fulfilling crises in the eurozone: An empirical test. *Journal of International Money and Finance* 34: 15–36.

De Grauwe, P., and C. Macchiarelli. 2015. Animal spirits and credit cycles. *Journal of Economic Dynamics and Control* 59(1): 95–117.

De Larosière Group. 2009. Report of 'The High Level Group on Financial Supervision in the EU'. Brussels, 25 February.

Eijffinger, S. 2013. The various roles of the ECB in the new EMU architecture. European Parliament Working Paper IP/A/ECON/NT/2013–03, part of the compilation PE 507.482 for the Monetary Dialogue.

European Central Bank. 2010. Recent developments in supervisory structures in the EU Member States (2007–10). October.

European Central Bank. 2011. The monetary policy of the ECB. Frankfurt.

European Central Bank. 2014. Financial integration in Europe, Chapter 2. April.

European Central Bank. 2015. Financial integration in Europe, Chapter 2 and Special Issue B. April.

European Commission. 2012. *A roadmap towards a banking union*. Communication from the Commission to the European Parliament and the Council, COM(2012) 510 final. Brussels: CEC.

European Parliament. 2012. The role of the ECB in financial assistance: Some early observations. EP Working Paper IP/A/ECON/NT/2012–04, PE 475.116.

Farhi, E., and J. Tirole. 2012. Collective moral hazard, maturity mismatch, and systemic bailouts. *American Economic Review* 102(1): 60–93.

Farhi, E. and Tirole, J. 2016. Deadly embrace: Sovereign and financial balance sheets doom loops. NBER Working Paper No. 21843, January.

Fuest, C. and Peichl, A. 2012. European fiscal union: What is it? Does it work? And are there really 'no alternatives'? CESifo Forum 1 (special issue on European Fiscal Union): 3–9.

Gabor, D. 2014. The ECB and the political economy of collateral. In *Central banking at a crossroads: Europe and beyond*, ed. C. Goodhart, D. Gabor, J. Vestergaard, and I. Ertürk. London: Anthem Press.

Gennaioli, N., A. Martin, and S. Rossi. 2014. Sovereign default, domestic banks and financial institutions. *Journal of Finance* 69(2): 819–866.

Gerba, E., and C. Macchiarelli 2015. Interaction between monetary policy and bank regulation: Theory and European practice. LSE Systemic Risk Centre Special Paper no. 10.

Goodhart, C.A.E. 2014. Lessons for monetary policy from the euro-area crisis. *Journal of Macroeconomics* 39: 378–386.

Goodhart, C.A.E., and D. Schoenmaker. 1995. Should the functions of monetary policy and banking supervision be separated? *Oxford Economic Papers* 47: 539–560.

Gros, D., and D. Schoenmaker. 2014. European deposit insurance and resolution in the banking union. *Journal of Common Market Studies* 52(3): 529–546.

International Monetary Fund. 2013a. IMF report on unconventional monetary policies–recent experience and prospects.

International Monetary Fund. 2013b. A banking union for the euro area. Staff Discussion Note, SDN/13/01.

Juncker, J.-C., D. Tusk, J. Dijsselbloem, M. Draghi, and M. Schulz. 2015. *Completing Europe's economic and monetary union*. Brussels: European Commission.

Macchiarelli, C. 2014. Banking union gaps leave European 'doom loop' intact, Oxford Analytica Daily Brief, September.

Masciandaro, D., R.V. Pansini, and M. Quintyn. 2013. The economic crisis: Did supervision architecture and governance matter? *Journal of Financial Stability* 9: 578–596.

Obstfeld, M. 2013. Finance at center stage: Some lessons of the euro crisis. European Economy, *Economic Papers* 493, Brussels.

Pisani-Ferry, J. 2012. The euro crisis and the new impossible trinity. In *Bruegel policy contribution 2012/01*. Bruegel: Brussels.

Posen, E. and Véron, N. 2014. Europe's half a banking union. Europe's World, 15 June. http://europesworld.org/2014/06/15/europes-half-a-banking-union/. Accessed 20 Oct 2016.

Reinhart, C.M., and K.S. Rogoff. 2013. Banking crises: An equal opportunity menace. *Journal of Banking and Finance* 37(11): 4557–4573.

Reinhart, C.M., J.F. Kirkegaard, and M.B. Sbrancia. 2011. Financial repression redux. *Finance and Development* 48: 22–26.

Schoenmaker, D., and N. Véron. 2016. European overview, Chapter 2. In *European banking supervision: The first eighteen months*, ed. D. Schoenmaker and N. Véron. Brussels: Bruegel.

Valiante, D. 2015. Banking union in a single currency area: Evidence on financial fragmentation. *Journal of Financial Economic Policy* 7(3): 251–274.

European Central Bank

Michael Binder and Volker Wieland

Abstract

The establishment of the European Central Bank (ECB) and with it the launch of the euro has arguably been a unique endeavour in economic history, representing an experiment of hitherto unknown magnitude in central banking. This article aims to describe the main aspects of the set-up and the responsibilities, strategy and operations of the ECB. It also aims to summarize some of the main lessons learned from the establishment of the ECB for monetary economics, and to sketch some of the prospects for the ECB and the euro.

Keywords

Budget deficits; Business cycles; Economic and Monetary Union (EMU); Euro; European Central Bank (ECB); European System of Central Banks (ESCB); Euro area; Exchange Rate Mechanism (ERM); Globalization; Harmonized Index of Consumer Prices (HIPC); Inflationary expectations; Maastricht criteria; Monetary transmission mechanism; Open market operations; Price stability; Reserve requirements; Standing facilities

JEL Classifications

E58; E5; F3; G1

The European Central Bank (ECB) was established on 1 June 1998 and since 1 January 1999 has been responsible for the conduct of a single monetary policy for its member countries, namely, Austria, Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain (with Greece subsequently becoming a member country on 1 January 2001 and Slovenia on 1 January 2007). Among European Union (EU) member countries Bulgaria, Czech Republic, Denmark, Estonia, Cyprus, Latvia, Lithuania, Hungary, Malta, Poland, Romania, Slovakia, Sweden and the United Kingdom are, as of 2007, not member countries of the ECB. These countries for the time being either have opted out of becoming member countries of the ECB (Denmark, Sweden and the United Kingdom) or have – according to the judgement of the EU Council – not yet achieved the necessary degree of economic convergence.

The launch of the ECB was the culmination of a process of monetary and economic integration that dates back at least to the efforts of the French government official Jean Monnet and others in the 1950s and gained decisive momentum with the April 1989 report of a committee headed by the then President of the European Commission, Jacques Delors, which drew up a blueprint for the progressive realization of the European Economic and Monetary Union (EMU). The establishment of the ECB and with it the launch of the euro (the currency of the ECB member countries for which banknotes and coins first went into circulation on 1 January 2002) has arguably been a unique endeavour in economic history, representing an experiment of hitherto unknown magnitude in central banking. In what follows, we shall describe the main aspects of the set-up and the responsibilities, strategy and operations of the ECB, discuss what appear to be the lessons learned from this experiment for monetary economics, and sketch some of the prospects for the ECB and the euro.

Lesson One: How to Converge?

There can be little doubt that the European Council's June 1989 decision to pursue the Delors Committee's blueprint of a feasible path towards monetary union for its member countries was primarily driven by political considerations, viewing monetary union as a building block towards tighter political and economic integration of the member countries of the EU. However, given the broad consensus among economists and policymakers that, ideally, economic similarity rather than political boundaries should define the

geographic area spanned by a common currency, the Delors report put considerable emphasis on realizing economic convergence before the establishment of a single European central bank. Key elements of the three stages to realization of the EMU as envisioned by the Delors Report were

- Stage 1 (1 July 1990): improvement of economic convergence; abolition of restrictions on cross-country flows of capital; increased cooperation between national central banks.
- Stage 2 (1 January 1994): strengthening of economic convergence; establishment of the European Monetary Institute (EMI) as predecessor of the ECB to strengthen cooperation between national central banks and increase coordination of monetary policy.
- Stage 3 (1 January 1999): completion of the necessary economic convergence; irrevocable fixing of currency conversion rates; single monetary policy to be conducted by the European System of Central Banks (ESCB).

It was envisioned in the Delors plan (and enacted in the Maastricht Treaty, which established the EU, as signed in February 1992) that only those countries should become member countries of the EMU that were successful in accomplishing economic convergence. The convergence criteria (Maastricht criteria) were meant to specify a sufficient degree of economic similarity of member countries with respect to price stability, sustainability of fiscal policy, exchange rate stability and the level of long-term interest rates. In particular, with respect to price stability member countries' average rate of inflation in the year preceding completion of the EMU was to fall within a one and a half per cent interval of average inflation in the three member countries displaying the highest degree of price stability. With respect to sustainability of fiscal policy, member countries were supposed not to carry an 'excessive deficit' – which would occur if the actual or planned government deficit to GDP ratio exceeded three per cent or if the ratio of government debt to GDP exceeded 60 per cent. Concerning exchange rate stability, member countries would in the two years preceding completion of the EMU have to keep the fluctuations of the value of their currency within the bands provided for by the European Exchange Rate Mechanism (ERM) and in particular not initiate any devaluation of their currency against that of any other member countries. Finally, with respect to the level of long-term interest rates, member countries' average long-term interest rates (on government bonds or comparable securities) in the year preceding completion of the EMU were to fall within a two per cent interval of average long-term interest rates in the three member countries displaying the highest degree of price stability.

Of course, economic similarities desirable for an optimal currency area do not end with these four criteria, but inter alia also include similarities in the monetary transmission mechanism, the coherence of the shocks and of the propagation mechanisms driving national business cycles as well as similarities in the prospects for trend output growth. These latter criteria were not part of the Maastricht criteria, though it was widely hoped that the economic convergence process prior to or immediately after the formation of the ECB would result in these latter criteria being approximately met as well.

Despite the relatively modest requirements for economic convergence in the Maastricht Treaty, the goal of EMU was jeopardized during the 1992–3 crisis of the ERM when foreign exchange market participants widely viewed the ERM's margins of fluctuation of two and a quarter per cent as not sustainable in the light of at best limited coordination of monetary policy, especially in Germany, with that in several other countries in the EU, specifically that in Italy and in the United Kingdom. The fact that despite the widening of the ERM's margins of fluctuation to 15 per cent in August 1993 the goal of EMU was maintained appears to have been due to the commitment of some of the then political leaders of the EU – perhaps most notably the then German Chancellor Helmut Kohl – who saw their vision of building a united Europe jeopardized. Owing to this political commitment as well as the fact that markets increasingly gave weight to complying with the Maastricht criteria as a signal for sound

monetary and fiscal policy, convergence as outlined by the Maastricht criteria was sufficiently advanced in May 1998 for the heads of state and government of the EU to decide to proceed with Stage 3 of EMU as planned, if only for the 11 initial member countries of the ECB.

While it is a valuable lesson to have observed in the context of the establishment of the ECB that the prospect of a monetary union may itself help to induce partial economic convergence, it appears key to keep in mind that the process of formation of the ECB would probably not have been successful without the strong desire of the member countries' political leadership to see commonalities in cultural heritage also reflected in increasingly cohesive institutional entities, trusting that a common European currency would help the emergence of a single European identity.

Structural economic diversities between euro area member countries continue today (in 2007). Among these diversities perhaps most notable are persistent differences in trend output growth rates. The widely voiced hope expressed at the time of the signing of the Maastricht Treaty – that formation of the ECB would significantly spur convergence of trend output growth rates for euro area member countries through alignment of structural reforms of labour and product markets – has so far proven to be wishful thinking. While some critics of the ECB have argued that this is due to the mandate of the ECB being too narrowly focused on price stability, it may have been exactly this focus that allowed the ECB to successfully establish itself as a credible safeguard of price stability, an issue which we will discuss further below.

Lesson Two: How to Design and Implement a Monetary Policy Strategy

The starting point for any discussion of the ECB's monetary policy strategy has to be the mandate that the ECB was given by the Maastricht Treaty. Article 105 of that treaty specifies: 'The primary objective of the ESCB is to maintain price stability. Without prejudice to the objective of price stability the ECB shall support the general economic policies in the Community with a view to contributing to the achievement of the objectives of the Community as laid down in Article 2.' Article 2 specifies these objectives to be a high level of employment as well as sustainable and noninflationary growth. (The Maastricht Treaty refers to the ESCB rather than the ECB since it envisioned that all member countries of the European Union would eventually adopt the euro and that even before this was to happen all national central banks of member countries not part of the euro area would be bound by the same objectives.)

While the Maastricht Treaty does not specify a precise quantitative definition of price stability, the ECB, particularly on the basis of the argument that such quantification would strengthen its commitment to its primary objective as well as strengthen its accountability, in October 1998 defined price stability as a year-onyear increase in the Harmonized Index of Consumer Prices (HICP) for the euro area of below two per cent over the medium run. While this definition of price stability does exclude deflation as being consistent with price stability and leaves the ECB with no degree of freedom to potentially remove more volatile and/or temporary components of overall consumer prices in order to declare price stability, the definition does leave the ECB some flexibility in that a time horizon as to what would constitute the medium run was not established.

In its pursuit of price stability, the ECB decided to base its monetary policy framework on two pillars: 'monetary analysis' and 'economic analysis'. In declaring monetary aggregates as providing information valuable to the objective of price stability that should be separated from other economic and financial variables, the ECB has so far maintained that monetary aggregates do not just offer incremental information relative to such other variables for purposes of projecting inflation, but that at longer horizons (stretching beyond those typically adopted by central banks for the computation of their inflation projections but still essential for medium-run price stability) monetary aggregates provide information qualitatively different from that which other economic variables can provide. The ECB in this context has so far also

maintained that money demand (as measured by the monetary aggregate M3) for the euro area has been stable at least over longer horizons, with some short-run instabilities being due to an exceptionally prolonged (but still temporary) period of high asset price volatility. Finally, the ECB has so far maintained that conventional macroeconomic analysis is not sufficiently advanced to combine the analysis of real economic phenomena with monetary trends within a single pillar framework. Driven by these considerations, the ECB therefore initially decided to announce annual reference values for the growth rate of M3 as a benchmark for keeping monetary growth in line with the objective of price stability.

The 'economic analysis' pillar of the ECB's monetary policy framework aims at identifying and quantifying short- to medium-term non-monetary risks to price stability. Variables entering this analysis include (*a*) gap measures of the discrepancy between actual output as well as its factors of production on the one hand and their medium- to long-run equilibrium values on the other hand; (*b*) labour cost measures; (*c*) exchange rates for the euro and international prices; and (*d*) asset prices other than exchange rates, particularly yield curve measures. Reflecting the sizeable degree of persistence of consumer price inflation in the euro area, considerable weight in the economic analysis is also given to recent consumer price dynamics.

The ECB's two-pillar strategy has been heavily criticized and remains controversial. Critics argue that monetary aggregates such as M3 – specifically due to the lack of sufficient stability of money demand – lack the degree of reliability needed to separate information in such monetary aggregates from other economic and financial variables. These critics inter alia also argue that, if the transparency and accountability of the ECB's decisions were to be improved, this would be helped most by the publication of inflation forecasts by the ECB as well as the publication of the minutes of the meetings of the ECB's Governing Council (for more on the latter, see below). The two-pillar strategy was reaffirmed in a broad internal assessment by the ECB in 2003, but two clarifications were provided. First, the Governing Council noted that it aims to maintain inflation rates below, but close to, two per cent over the medium run. A number of arguments in favour of tolerating a low rate of inflation – and not aiming at zero inflation – were acknowledged, among which the most important are the need for a safety margin against potential risks of deflation and the 'zero bound' on nominal interest rates. While this 'zero bound' renders central bank interest-rate management less effective at low rates of inflation, ECB studies argued that inflation rates below, but close to, two per cent would provide a sufficient safeguard against these risks. Second, the Governing Council emphasized that the 'monetary analysis' pillar was meant to serve mainly as a means of cross-checking, from a medium- to long-term perspective, the short- to mediumterm indications provided by the 'economic analysis' pillar. To underscore the longer-term nature of the reference value for monetary growth, the practice of an annual review of the latter was discontinued.

It will be interesting to observe whether eventually the monetary pillar comes to be viewed as having been of importance only in the early years of operation of the ECB when the ECB had to establish its credibility by being as committed to price stability as the Deutsche Bundesbank (the German central bank) had been prior to 1999 and when the ECB was confronted with sizeable problems regarding the measurement of harmonized euro area-wide real economic aggregates, or whether ECB-style cross-checking by means of monetary analysis will become a common practice of central banks around the globe.

The operational framework used by the ECB to implement its monetary policy strategy is less controversial than the strategy itself and includes three main instruments: open market operations, standing facilities and reserve requirements. Among the open market operations of primary importance are the 'main refinancing options' that provide the bulk of refinancing to the financial sector and, through signalling the ECB's monetary policy stance, are supposed to steer market interest rates. The 'main refinancing options' are executed by the national central banks of the euro area member countries on a weekly basis through

a tender procedure spanning three working days. 'Standing facilities' aim at providing and absorbing overnight liquidity, and 'minimum reserve requirements' (the ECB imposes minimum reserves on all credit institutions in the proportion of two per cent of the reserve base) aim at stabilizing market interest rates.

By way of evaluating the overall success of the ECB in terms of it being able to adhere to its price stability objective, we may observe that inflation rates in the euro area since 1999 have on an annual basis on average been slightly above two per cent (in the range of up to 30 basis points above two per cent). Also, given that surveys of average long-term inflation expectations in the euro area have consistently measured such expectations as below, but close to, two per cent, its track record has quite firmly established the ECB's credibility with regard to safeguarding price stability.

Lesson Three: One Central Bank for Many Countries: How to Organize Decision-Making

The most important decision-making body of the ECB is its Governing Council, which is made up of the Executive Board of the ECB (which in turn is made up of its president, vice-president and four other members) as well as the governors of all the national central banks of euro area member countries. It is the responsibility of the Governing Council to formulate monetary policy for the euro area, including decisions about intermediate objectives and key interest rates. The Executive Board is in charge of implementing the monetary policy decisions taken by the Governing Council, and to this purpose cooperates with the national central banks through open market activities. Each member of the Governing Council has one vote. Given that at present slightly more than two-thirds of the votes in the Governing Council, therefore, belong to national central banks, the latter have a strong influence on the ECB's monetary policy decisions.

This organizational structure implies an asymmetry between the economic size of euro area member countries and their influence on decisions arrived at by the Governing Council. Indeed, more than half the euro area member countries at present have an economic weight (as measured by the ratio of their national GDP to euro area GDP) that is smaller than their voting weight within the Governing Council. This is quite different from the structure of, say, the US. Federal Reserve, which is significantly more centralized. While decentralization of the implementation of the ECB's monetary policy arguably is useful, particularly as long as there are important differences among national financial markets and institutions in the euro area, the decentralized institutional set-up of the ECB has risks, particularly during episodes of real divergence. It will be interesting to see whether the 'one person, one vote' principle for the Governing Council will be maintained after possible enlargement of the euro area to incorporate (some of) the EU member countries not presently member countries of the ECB. Even if the 'one person, one vote' principle is to be maintained, there appears to be considerable scope for future revision of the organizational system of the ECB, such as requiring approval of nominations of new central bank presidents by the Executive Board of the ECB.

Lesson Four: Common Currency and Monetary Policy: Gains and Losses

In general, the principal advantages of a common currency are widely held to include the reduction of transaction and information costs implied by the use of a common medium of exchange as well as the stimulus the common currency provides for the convergence of organizational principles used in business, in turn stimulating trade in goods and services and of cross-country flows of capital. The principal disadvantages of a common currency for multiple countries are widely held to include the loss of shock-absorber properties of flexible exchange rates and of independent national monetary policies. Furthermore, if a single monetary policy is accompanied by a diverse set of national fiscal policies, inappropriate fiscal policy in one country will – through its effect on interest rates – directly spread to other countries in the

monetary union. Thus macroeconomic stability could be affected for the worse.

How has the euro area so far fared on these counts? Trade within the euro area increased from approximately 26.5 per cent of (euro area) GDP in 1998 to approximately 31 per cent of GDP in 2005; one and a half per cent of this increase was due to trade in services. Taking into account the limited time span, it is difficult to assess, however, to what extent this increase in trade was indeed driven by the creation of a single currency and to what extent it may instead have been driven by the process of economic globalization. We do know, in fact, that trade with trading partners outside the euro area over this same time period rose by a slightly larger margin than intra-euro area trade, from approximately 24 per cent of GDP in 1998 to approximately 30 per cent of GDP in 2005.

Regarding financial markets, for which the volume of transactions is probably still more sensitive to even small costs and risks associated with the use of multiple currencies, by a variety of measures deeper, broader and more liquid markets have emerged for the euro area member countries since establishment of the ECB. On the money market, issues of their interpretation aside, cross-country standard deviations for average overnight lending rates fell from 130 basis points in January 1998 to three basis points one year later, and since then have decreased to approximately one basis point. Cross-country standard deviations for rates at longer maturities (one and 12 months) for unsecured money market instruments have fallen to less than one basis point also, with the spreads still somewhat larger in the collateralized repurchase agreement (repo) market (due to continued differences in legal structures across euro area countries). In the interest rate derivatives market, the euro interest rate swap market at a daily volume of 250 billion euro was in 2006 one and a half times as large as the corresponding US dollar market. In the government bond market also, spreads have fallen to low levels, suggesting – in the likely absence of major changes in default risks – a significant fall of liquidity risk. The holdings of euro-denominated debt securities overall since 1999 have increased by well over ten per cent to approximately one-third of the global market (through holdings tend to be concentrated in countries neighbouring the euro area).

In the equity and retail banking markets integration has progressed more slowly. For example, despite a decrease in the number of credit institutions in the euro area member countries by almost 50 per cent between 1997 and 2006, less than one-third of the mergers and acquisitions driving this consolidation process have been crossborder. Also, the cross-country standard deviation of interest rates on consumer credit from 2004 to 2006 has still been close to one per cent.

While, just as for trade, it is difficult to disentangle the euro's contribution to the process of financial integration in euro area member countries from the global trend towards financial integration, the euro surely has greatly facilitated the task of bringing the European financial system closer to US standards in terms of market depth and liquidity. Further improvements in this direction, including the creation of a single payment system for the euro area member countries, are likely to intensify the debate about the potential role of the euro as a complement or competitor to the US dollar as an international reserve currency.

Finally, to turn to macroeconomic stability and the potential cost of losing flexible exchange rates and independent national monetary policies as shock absorbers, some such costs clearly have been observed since 1999. While the cross- country standard deviation of consumer price changes has fallen from approximately six per cent in the late 1990s to one per cent with the launch of the euro, and has been rather stable at this level in the following eight years, there have been persistent deviations from euro area average inflation rates for some countries, implying sizeable (and potentially destabilizing) differences in real interest rates. For example, for a sizeable part of the time period since 1999, real interest rates have been significantly lower in a booming Irish economy than in a German economy experiencing weak growth. When it comes to assessing the implications of the establishment of the ECB for macroeconomic stability, these costs have to be subtracted from benefits owed to factors such as

the elimination of intra-euro area exchange rate crises and the fact that inflation rates for some euro area member countries have been falling sizeably in the eight years since 1999. However, a stronger degree of real convergence through aligned policies aimed at removing structural deficiencies in European product and labour markets would have helped to render the benefits yet larger.

Conclusion

While this article has suggested that on various counts (such as the monetary policy strategy and the organizational set-up) there is as of 2007 no consensus as to whether the ECB adheres to best international practice in central banking, it would appear rather questionable to label the establishment of the ECB and with it the introduction of the euro as anything but an enormous success. The ECB has successfully mastered the technical challenges of establishing a new common currency across a set of countries comprising one of the largest economic regions in the world, has in a short period of time established a strong track record of success in preserving price stability, and has on many counts, particularly in the area of financial markets, helped lead the way to a stronger integration of European markets. While it is undisputable that this integration of markets along with structural reforms needs to proceed much further, the key decisions that could facilitate such integration and structural reforms fall outside the core domain of responsibility of the ECB and, for that matter, should probably remain so for any central bank primarily entrusted with maintaining price stability.

See Also

- ▶ Euro
- ▶ European Monetary Union
- ▶ Federal Reserve System
- ▶ Inflation Targeting
- ▶ Stability and Growth Pact

Bibliography

European Central Bank. 2004. *The monetary policy of the ECB*. ECB: Frankfurt am Main.

European Central Bank. 2006. *ECB statistical data warehouse*. Online. Available at http://sdw.ecb.int. Accessed 14 Mar 2007.

Issing, O. 2003. *Background studies for the ECB's evaluation of its monetary policy strategy*. ECB: Frankfurt am Main.

Issing, O., V. Gaspar, I. Angeloni, and O. Tristani. 2001. *Monetary policy in the euro area*. Cambridge: Cambridge University Press.

Padoa-Schioppa, T. 2004. *The euro and its central bank: Getting united after the union*. Cambridge: MIT Press.

Posen, A.S. 2005. *The euro at five: Ready for a global role?* Washington: Institute for International Economics.

European Central Bank and Monetary Policy in the Euro Area

Vítor Gaspar and Otmar Issing

Abstract

Since 1 January 1999 the European Central Bank (ECB) has had sole responsibility for monetary policy in the euro area. Its main aim is to maintain price stability over the medium term. It is completely independent. The ECB has been successful in maintaining price stability and well-anchored inflation expectations. The euro is now a well-established international currency and a symbol of European integration. The financial and economic crises of 2007–2011 made the tension between the single monetary policy and national responsibilities for economic policies and financial stability visible. This is the main challenge going forward.

Keywords

Euro; European Union; European Central Bank; Monetary policy; Price stability; Expectations; Credibility; Finance

JEL Classifications

E52; E58; E61; F55

A single monetary policy covering all the sovereign states participating in the euro area replaced the separate national policies on 1 January 1999. The date is also a milestone in the history of international monetary integration. For the first time a group of advanced countries has chosen to entrust the exclusive competence for the conduct of their monetary policy to an independent and supra-national monetary authority: the European Central Bank (article 3 and article 127 of the Treaty on the Functioning of the European Union – EU). The euro area started with the participation of 11 EU Member States: Austria, Belgium, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Portugal and Spain. Since then, it has expanded to include 17 countries with the participation of Greece (2001); Slovenia (2007), Cyprus and Malta (2008), Slovakia (2009) and Estonia (2011). Euro banknotes and coin became a physical, day-to-day reality on 1 January 2002.

At the time of writing (2011), the euro is the currency of almost 330 million people in 17 different European countries. Most of the remaining EU Member States intend to adopt the euro in the future and, with the exception of the UK and Denmark, have a treaty obligation to do so. The euro is the second most important currency in the global economy (after the US dollar) and the euro area is the monetary area with the second largest economy (after the USA). In contrast with the economies of the participating countries, which can mostly be characterized as small open economies, the euro area is a large and relatively closed economy. Aggregate differences, in the structure of production by sector, relative to the USA are relatively small. For the conduct of monetary policy, an important difference relative to the USA is that in the euro area the financial system is dominated by banking. This contrasts with the predominance of market financing in the USA. For more information on the global role of the euro, the economic weight of the euro area and further references see, for example, EMI (1997c), ECB (1999), Issing et al. (2001) and Buti et al. (2010).

The transfer of powers from the participating Member States was limited to monetary policy. Their competences in economic policy, in general, and budgetary policy, in particular, remained unchanged. The combination of an independent and supranational monetary authority, with Member States largely responsible for the conduct of economic policies, implies an original tension that remains unsolved. Many observers were sceptical about the outcome. After a few years the pendulum swung the other way and most observers started taking the success of the euro for granted, forgetting how difficult it had been to prepare the launch of the new currency, to establish the credibility of the ECB and to put in place a new monetary policy strategy and operational framework. The degree of uncertainty associated with the transition to the single monetary policy was enormous. Issing (1999b) writes:

> As a central banker directly involved in monetary policy making, I have been dealing with uncertainty and its consequences for a large part of my professional life. From my experience as a member of the Board of the Bundesbank, I have vivid memories of challenges posed by German reunification and the turbulence surrounding ERM crises. But never have I felt the impact of uncertainty as acutely as in the weeks that preceded and followed the introduction of the euro and the birth of the single monetary policy.

According to the Treaty on the Functioning of the European Union the primary objective of the European System of Central Banks (ESCB) is to maintain price stability (article 127.1 and 282.2, and also article 2 of the Statute of the European System of Central Banks and of the European Central Bank). The ESCB is governed by the decision-making bodies of the ECB (article 129.1 and 282.2 of the Treaty and 9.3 of the Statute): the Governing Council and the Executive Board. The independence of the ECB and of the ESCB is protected by the Treaty and the Statute (respectively by article 130 and by article 7). Without prejudice to the objective of price stability, the ESCB shall support the general economic policies in the Union with a view to achieving their objectives, as specified in article 3 of the Treaty on European Union. These include:

- balanced growth;
- a highly competitive social market economy, aiming at full employment and social progress;

- the promotion of scientific and technological advance;
- equality between women and men;
- solidarity between generations;
- the protection of the rights of the child;
- economic, social and territorial cohesion; and
- solidarity among Member States.

The ESCB is composed of the ECB and the national central banks of the Member States of the European Union. The Eurosystem is made up of the ECB and the national central banks of the Member States that have adopted the euro as their currency. These definitions are provided in article 1 of the Statute.

Most of the preparatory technical work was carried out by the European Monetary Institute (EMI) and the participant national central banks during the second stage of Economic and Monetary Union (1994–1998). During this period the foundations for the proper functioning of the single currency were conceived. These included a basic set of analytical tools; statistical information; internal organization rules for the new central bank; a new pan-European interbank payment mechanism; and the operational framework for implementation of the single monetary policy. The technical work was already mature when the ECB was established (on 1 June 1998) and the final six months of the preparatory period started (see, for example, EMI 1997a, b, 1998).

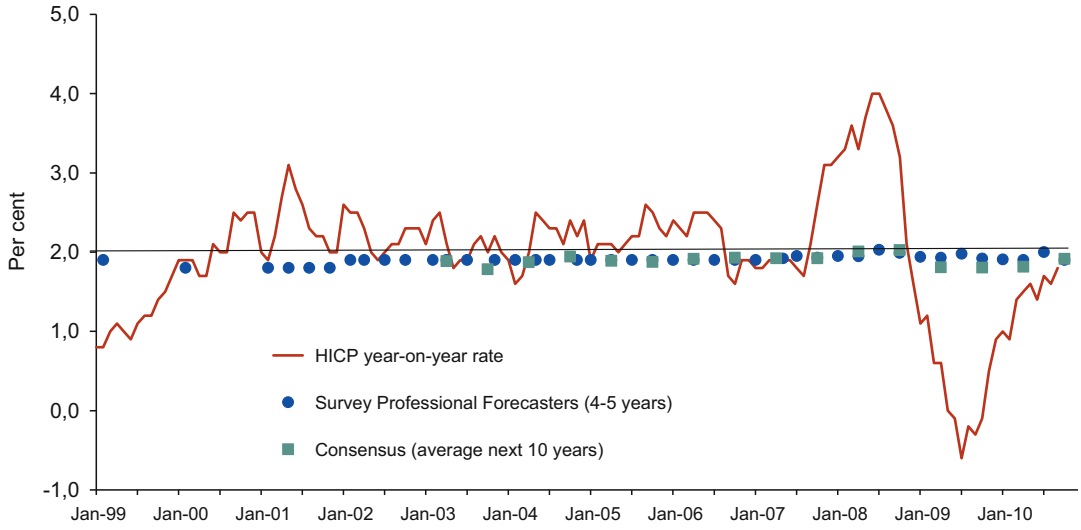
The institutional provisions described in the previous paragraphs underline the importance of price stability as a central element in the economic constitution of monetary union in Europe. However, all institutional guarantees notwithstanding the most basic questions at the beginning were: would the ECB deliver price stability? Would the ECB be credible in fostering well-anchored inflation expectations? How could the ECB be credible in the absence of a track record? How could the ECB deal with the uncertainties associated with monetary unification and financial integration?

The importance of credibility for the conduct of monetary policy is a common feature of virtually all models that emphasize the endogenous character of private sector expectations. Expectations management is strongly emphasized in the academic literature, starting with Kydland and Prescott (1977) and Barro and Gordon (1983). This important issue has been revisited recently, in the context of the standard new Keynesian model, by, for example, Clarida et al. (1999), Woodford (2003), Galí (2008) and Walsh (2010). This literature stresses that an independent central bank, with a clear mandate, should be able to maintain low and stable inflation and to anchor inflation expectations. Credible policy relies on a well-understood strategy, implying a systematic and predictable pattern of response to the current state and prospects for the economy.

Figure 1 shows inflation in the euro area, measured in accordance with the 12 month change in the Harmonized Index of Consumer Prices (HICP) and long-term inflation expectations following the Consensus forecasts and the ECB's own Survey of Professional Forecasters (SPF). In the period from January 1999 to December 2010, average inflation was 1.97%, which is below (but close) to 2% and therefore in line with the ECB's definition of price stability. Importantly, the same applies to the long-run inflation forecasts according to the Consensus forecasts and the SPF. For example, according to the latter average, long-term inflation expectations have always remained well-anchored within a narrow range from 1.8% to 2.0%.

Figure 1 also makes it clear that year-on-year inflation has been above 2% most of the time. In June and July 2008 it peaked briefly at 4% (twice the upper limit in the ECB's definition of price stability). Equally, for most of the period, average annual inflation has been close to but not below 2%. Only recently, in the context of the Global Financial Crisis and the associated Great Recession, have very low – or even, for a brief period, negative – inflation rates pushed average inflation below the 2% limit.

In Buti et al. (2010), Geraats, Neumann and Smets look back at the performance of the ECB during the first decade of the euro. They emphasize the behaviour of inflation expectations and credibility. For example, Smets (2010) looks at the first decade of the euro. He excludes the first year (1999) on the ground that given transmission lags the outcome cannot be attributed to the ECB. Hence, for the period available to him, HICP



Sources: Consensus Economics, ECB and Eurostat.

European Central Bank and Monetary Policy in the Euro Area, Fig. 1 Euro area – inflation (HICP) and long-term inflation expectations

inflation, for the euro area, averaged 2.2%. Smets argues that the deviation is mostly due to unforeseen and large oil and other commodity prices disturbances. In fact, excluding energy and unprocessed food prices, the average inflation rate was 1.8% in the same period. But mostly he stresses, as already mentioned above, that while headline inflation has fluctuated significantly long-term inflation expectations remained anchored, in line with the ECB's definition of price stability.

European monetary unification has been debated since the establishment of the so-called Werner Group (1969). The discussion intensified after the publication of the Delors Report in 1989. In the academic world, the vast majority of commentators have been highly sceptical and critical (see Issing (2008) and Jonung and Drea (2010) for reviews and references). Even before the start of the single monetary policy several groups of economists organized ECB watching activities. The title of the first of these reports, CEPR's 'ECB: safe at any speed?', published in 1998, is representative of the prevailing tone at the time. Since 1999 the ECB has been meeting annually with academics and professionals from the financial sectors at 'ECB and Its Watchers Conferences'. Thee meetings are organized by the Center for Financial Studies (CFS) at Frankfurt University. These conferences, unique in the world of central banking, provide a very full picture of the ongoing debate between the ECB and its critics (the programme of the first 12 editions of the conference and a wealth of additional information can be obtained from the CFS website: http://www.ifk-cfs.de/index.php?id=1164). Whereas the monetary policy decisions were mostly welcomed, the initial criticism concentrated on the ECB's monetary policy strategy. However, the critical tone lessened somewhat over time. In the context of the financial crisis, even early critics have seen the ECB less critically than other major central banks (see, for example, Buiter (2009)).

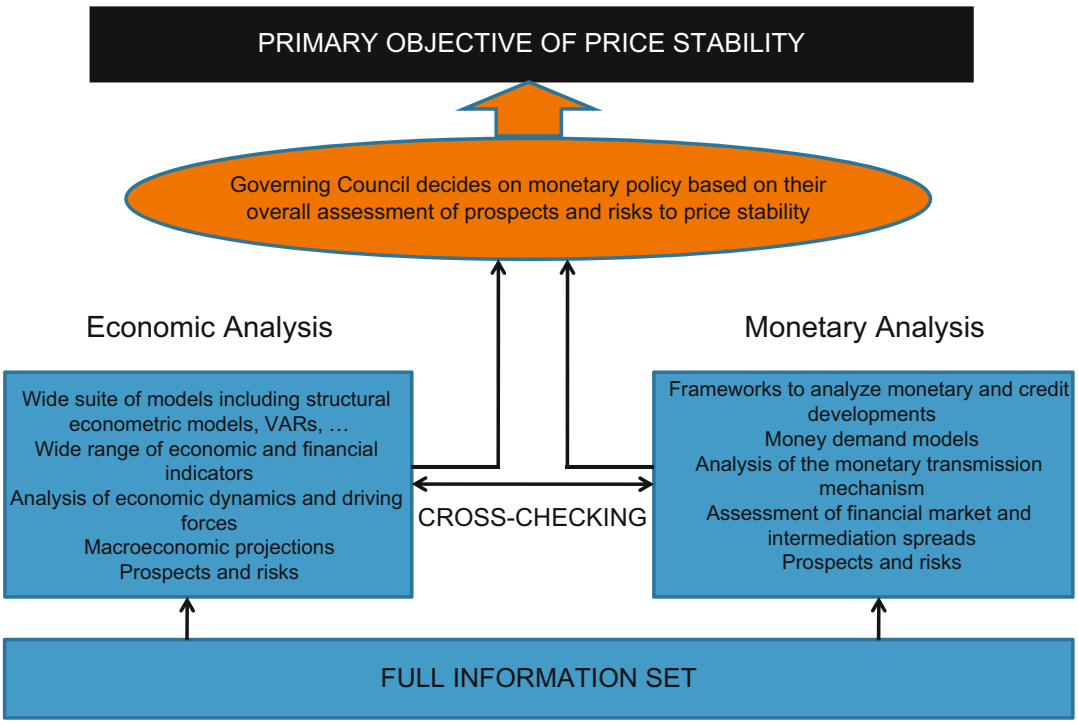
The Stability-Oriented Monetary Policy Strategy of the ECB

The environment relevant for the conduct of monetary policy is characterized by pervasive uncertainty. As argued above, this general point was of particular relevance for the ECB at the time of the launch of the euro. Moving from national currencies and monetary policies to a common currency and a single monetary policy implied a deep

regime shift with a huge potential for structural breaks (Lucas 1976). However, knowledge about the structure and the functioning of the economy and of the financial system is always imperfect. The structure and functioning themselves are constantly changing, which limits the usefulness of past data. Therefore it is the case that the monetary transmission mechanism is always uncertain. Moreover, economic data are contaminated by measurement error. In this environment it is crucial that monetary policy does not itself become an additional source of uncertainty. A monetary policy strategy helps to dispel such uncertainty, first by structuring a well-ordered internal decision-making process; second, by providing a consistent and coherent framework for communication; and third, by contributing to the credibility and predictability of the single monetary policy.

The stability-oriented monetary policy strategy of the ECB was disclosed immediately after a decision by the Governing Council, on 13 October 1998 (ECB 1998). The strategy comprises three elements: (1) a quantitative definition of price stability: (2) economic analysis and (3) monetary analysis. The latter two are used to organize an all-encompassing assessment of price prospects and risks to price stability. The two perspectives are systematically used for cross-checking (see Fig. 2). Versions of Fig. 2, differing in some details, have been used over time by the ECB to convey a schematic representation of the monetary policy strategy. It was presented for the first time in an *ECB Monthly Bulletin* article in 2000 (ECB 2000).

It is important to clarify the status of the definition of price stability within the monetary policy strategy. In fact, as already referred to above, price stability is set as the primary goal of the ECB in the Treaty on the Functioning of the European Union itself and the ECB does not, as some critics have suggested, have independence in setting its goal. Rather, price stability as the overriding goal of monetary policy is taken as given by the ECB (Issing 2000). Instead, the ECB's decision was to



European Central Bank and Monetary Policy in the Euro Area, Fig. 2 The stability-oriented monetary policy strategy of the ECB

announce a precise and operational definition of price stability based on a specific statistical indicator. Such an announcement is, in itself, an important form of commitment and a key for the communication of the central bank with the general public (Issing et al. 2001, Chapter 4).

In October 1998, the Governing Council decided that 'Price stability shall be defined as a year-on-year increase in the HICP, for the euro area, of less than 2 per cent. Price stability is to be maintained over the medium term'. About five years later, on 8 May 2003, when announcing the results of its evaluation of the monetary policy strategy, the Governing Council confirmed the definition, but clarified that it aimed to maintain inflation below (but close) to 2% over the medium term. The clarification was justified by the benefits of making explicit a safety margin against the risk of deflation. The Governing Council spelled out that it regards low and stable inflation as compatible with price stability. Inflationary and deflationary departures from the benchmark are both undesirable departures from price stability. Therefore the definition is clearly symmetrical.

The definition of price stability provides a benchmark against which to evaluate the performance of the ECB. It provides an anchor for inflation expectations and therefore serves to reduce uncertainty about price developments over the longer term.

Pervasive uncertainty also recommends a focus on robustness. Seeking robustness involves a willingness to consider different 'views of the world'. In the ECB's stability-oriented monetary policy strategy such diversity is symbolized by 'two pillars', supporting decision-making by the Governing Council: economic analysis and monetary analysis (Fig. 2).

Economic analysis spans a wide range of indicators which are relevant for risks to price stability over the short to medium term. This includes overall output, demand and labour market conditions, fiscal policy, and exchange rate developments, as well as financial market indicators and asset prices. Economic projections are an important element of economic analysis as they synthesize a very rich information set. They do not, however, constitute a sufficient statistic. The projections are produced by staff, as an input to Governing Council deliberations, at a quarterly frequency. Twice a year they stem from a broad exercise involving not only the ECB but also the national central banks of the Eurosystem. Twice a year they are conducted under the sole responsibility of ECB staff. Economic analysis focuses, to a large extent, on the interaction between aggregate supply and aggregate expenditure and the role of factor costs on pricing behaviour. Economic analysis uses a vast array of structural econometric models, including new Keynesian Dynamic Stochastic General Equilibrium Models. The ECB has been pioneer in this area of research through the work of Smets and Wouters (2002). At the time of writing the new area-wide model (NAWM), developed by the Econometric Modeling Division, to be used in the broad macroeconomic projection exercises and policy simulations, constitutes an important element in the ECB's modelling toolbox. It is a micro-founded, open-economy model for the euro area. It relies on a neo-classical core and incorporates a number of important frictions including wage and price rigidities; habit persistence in consumption behaviour; and adjustment costs in investment. It incorporates some open economy extensions of these frictions, including domestic currency pricing and costs associated with the adjustment of trade flows (Christoffel et al. 2008). The NWAM follows the area-wide model (AWM) that was developed and made available in the early years of the ECB (Fagan et al. 2001).

Monetary analysis starts from the fundamental insight that inflation is ultimately a monetary phenomenon (according to Milton Friedman). Inflation, that is persistent increases in the price level, is ultimately determined by monetary trends. Although many factors may affect price behaviour in the short to medium term, only monetary trends can account for lasting inflation (see, for example, Romer 2006, p. 407). The money–price relationship is confirmed by a wide variety of empirical studies using times series, cross-country and pooled data, spanning different monetary regimes and definitions of monetary aggregates. A central bank with the mandate to maintain price stability cannot ignore its responsibility for monetary developments. By giving 'money' a prominent

role, in its monetary policy strategy, the ECB has recognized this role and, at the same time, avoided some of the shortcomings of inflation targeting. Therefore monetary analysis contributes to cross-checking economic analysis from a long-term perspective. Monetary analysis may also be relevant at shorter horizons through, for example, the monitoring of credit developments, financial spreads and the monetary transmission mechanism. More generally, cross-checking the results from economic and monetary analysis provides the foundation for monetary policy decisions, which take into account all relevant information, seen from two different angles (see Fig. 2).

Beck and Wieland (2008) show that, when the central bank misperceives the output gap, supplementing interest rate prescriptions, derived from Keynesian or new Keynesian models, with estimates of trend inflation derived from monetary analysis, substantially improves inflation outcomes.

From the beginning monetary analysis was not restricted to broad definitions of money such as M3 and its relation to the reference value, but took into account 'developments of a wide range of monetary indicators, including M3 and its components and counterparts, notably credit and various measures of excess liquidity'(ECB 2003a). Over time, monetary analysis was broadened and deepened (Issing 2005a).

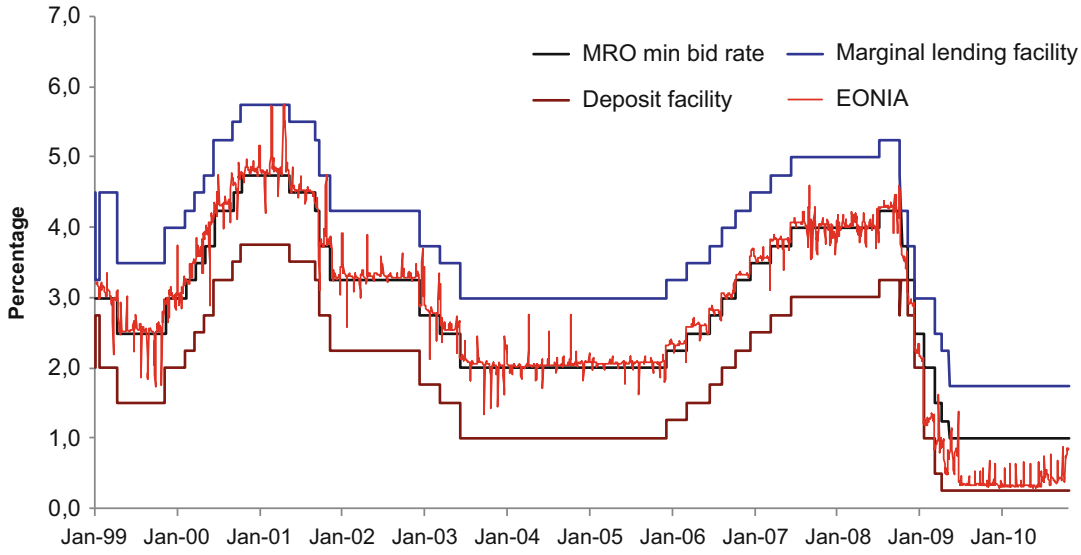
In the context of the Global Financial Crisis, starting in 2007, links between liquidity and asset price dynamics have become evident, stressing the relevance of monetary analysis also for identifying risks to financial stability. There is much ongoing research at the ECB and elsewhere. Recently Beyer (2009) has estimated an empirically stable, small macroeconomic model for the euro area. It considers the broad monetary aggregate, M3, of real GDP, annual inflation, the nominal growth of housing wealth and interest rate measures (specifically the annualized three-month interest rate and the annualized own return on M3). The model is able to track closely trend velocity since the late 1990s. While the money demand model identifies the influence of wealth variables on money demand another strand of research, building on insights from Hyman Minsky (see, for example, Minsky 1975), looks at the reverse link. Money and, in particular, credit growth in excess of what is needed for sustainable growth begets asset price bubbles and financial instability. Therefore they are associated with boom and bust and, from a medium-to long-term perspective, with price instability. Alessi and Detken (2009) have found that the global private credit gap is a leading indicator of asset price booms that will be followed by costly episodes of macroeconomic instability and ultimately price instability.

A volume recently published by the ECB (Papademos and Stark 2010) presents studies on various aspects confirming the importance of a thorough monetary analysis while providing subjects for further research (summaries are provided in ECB (2010) and Amisano et al. (2010)).

The monetary policy strategy of the ECB was designed in a way that, notwithstanding basic principles, is open to new insights from research and actively seeks to identify structural changes in the financial system as well as in the real economy. This approach is confirmed by this statement (Papademos and Stark 2010, p. 11): 'Developing a better understanding of the behaviour of monetary and credit aggregates and their influence on the economy has given rise to new questions and challenges which will influence our future work on both monetary analysis and other approaches. Reciprocally, improvements in other approaches will again need to be considered from the perspective of monetary analysis. Both to crosscheck their results and to maintain the encompassing nature of a robust monetary policy strategy. Such is the nature of progress. Such is also the nature of science. And such should be the nature of a robust monetary policy strategy if it is to ensure effectiveness, accountability and transparency'.

The Operational Framework for Monetary Policy Implementation

The Eurosystem implements monetary policy through financial markets. The operational framework is the set of instruments and procedures through which the Eurosystem intervenes in markets with a view to determining the monetary



Source: Thomson Reuters

European Central Bank and Monetary Policy in the Euro Area, Fig. 3 ECB policy rates and EONIA

policy stance, in accordance with decisions made by the Governing Council. The Treaty and the Statute contain relatively few provisions about the instruments of monetary policy. Furthermore, those provisions are of a general nature. For example, article 127.1 of the Treaty on the Functioning of the European Union (article 2 of the Statute) prescribes that the Eurosystem 'shall act in accordance with the principle of an open market economy with free competition favoring an efficient allocation of resources'. The operations of the Eurosystem are specified in chapter IV of the Statute (articles 17–24). For example, the Eurosystem may operate in financial markets (article 18), and impose minimum reserves on credit institution, including penalties for non-compliance (article 19). Importantly, article 20 states: 'The Governing Council may, by a majority of two thirds of the votes cast, decide upon the use of such other operational methods of monetary control as it sees fit. . .'. Therefore, the Treaty and the Statute provides the ECB with ample room to adapt its instruments as required by unforeseen circumstances. Naturally, discretion in this area is, as in all other areas, constrained by the primary goal of maintaining price stability, as prescribed by article 2.

Under normal circumstances, the first step in the monetary transmission mechanism is the control, by the central bank, of an overnight interest rate. The ECB does so through a 'corridor system' (see Fig. 3). In a 'corridor system', reserve requirements operate as a device to create a structural demand for central bank money and a buffer to smooth demand for reserves and reduce the volatility of interest rates. Overnight interest rates (in Fig. 3 represented by the Euro Overnight Index Average (EONIA)) are bound by two standing facilities provided by the central bank: a deposit facility and a lending facility. Banks can deposit excess funds at a predetermined interest rate. Banks can also have recourse to a credit facility, accessing funds at a predetermined interest rate, securing the operation through the pledging of eligible collateral (see 'Marginal lending facility' in Fig. 3). Within the corridor, the Main Refinancing Operation (MRO) minimum bid rate is the most important policy rate, as it serves as the reference for overnight market interest rates (for details see Bindseil (2004) and ECB (2008)). Issing (2008) offers an overview of the process of development of monetary policy instruments by the Eurosystem).

ECB Policy in Action: Simple Rules as Benchmarks

In the models used by monetary policy researchers, the rules summarizing the systematic response of interest rates to the state of the economy are central. In forward-looking models under perfect foresight or rational expectations, the model cannot be solved in the absence of such rules. Focus on systematic (rule-like) behaviour is not, however, the most common way of discussing the practice of monetary policy. Many people fail to distinguish between actions taken in the context of a systematic practice and the characterization of the features of the practice itself. Nevertheless, it is a fundamental distinction. The stability-oriented monetary policy strategy of the ECB induces systematic, rules-like behaviour on the part of the ECB. It is therefore interesting to assess how closely simple rules are able to account for the past behaviour of the ECB. Close tracking is direct evidence of systematic behaviour. Significant departures are episodes that demand explanation and clarification on the basis of the strategy itself.

In normal times, the Eurosystem conducts monetary policy through control over money market interest rates. Therefore the most straightforward example of a simple policy instrument rule is to express the interest rate as a function of a small set of relevant variables. Review reveals relatively little evidence that hosting the Games produces significant economic benefits for the host cityTaylor rule (Taylor 1993) that expresses the interest rate as:

$$i = r^* + \pi + \theta_\pi + (\pi - \pi^*) + \theta_y(y - y^*)$$

where the interest rate is i; r^* is the natural rate of interest; π is the rate of inflation; π^* is the inflation rate deemed compatible with price stability and y^* is the level of potential output (in logs). Taylor originally proposed $r^* = 2.0$ and $\theta_\pi = \theta_y = 0.5$.

In empirical studies it is common to consider also a term with the lagged interest rate on the right-hand side so as to reflect interest rate smoothing. Another simple rule has been proposed by Orphanides (2003, 2006, 2010). It has the form:

$$\Delta i = \theta[(\pi - \pi^*) + (\Delta y - \Delta y^*)]$$

The Orphanides rule departs from the Taylor rule in that, by using first-differences, it eliminates the dependence of the rule on the unobserved level of the natural interest rate, r^*, and also on the level of potential output, y^*. Following in Taylor's footsteps, Orphanides proposes $\theta = 0.5$. Smets estimates the rule and finds that the estimated coefficients on inflation and output growth deviations are not significantly different from 0.5. The coefficient on lagged interest rate is close to 1 (but the estimated coefficient is 0.89 and the difference from 1 is statistically significant).

The Orphanides rule may be derived from a quantity theory of money framework (Orphanides 2003). Indeed, the ECB derives the reference value for money growth from the equation of exchange:

$$\Delta m^* = \pi^* + \Delta y^* + \Delta v^*$$

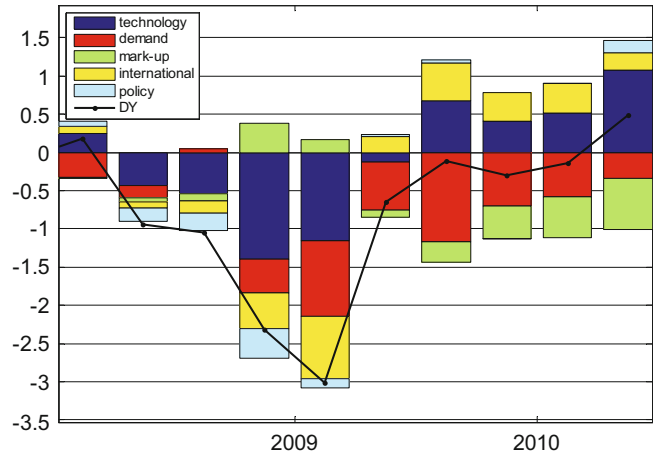
as explained in Issing (2008). In order to derive the interest rate rule it suffices to use the simplest money demand equation – which relates deviations of velocity from long run trends to the relevant interest rate and transitory departures from trend – and to ignore temporary disturbances.

The NAWM includes a Taylor-type monetary policy reaction function that includes interest-rate smoothing, the deviation of aggregate output from the trend implied by permanent technology shocks as the measure of the output gap and an inflation objective that may temporarily deviate from its long run value.

Issing et al. (2001, p. 42) state:

> Simple rules, in particular of the Taylor type, appear to provide *ex post* a good description of policies actually followed by central banks in the eighties and the nineties (though with notable exceptions such as the period of ERM crisis in Europe).... In spite of these good descriptive properties, simple rules are rarely advocated as *prescriptive* policy tools, even by their proponents. A total commitment to a simple rule could lead to sub-optimal policies since, by assumption, they do not take into account all potential sorts of information that can, from time to time, be relevant for monetary policy (for example, financial crisis or asset market bubbles).

E



European Central Bank and Monetary Policy in the Euro Area, Fig. 4 Decomposition of output growth into the shocks (*Source*: Christoffel et al. (2008), updated by courtesy of the authors of the NAWM)

The simple difference rule, proposed by Orphanides, and the Taylor rule used in the NAWM are two examples (among many possible) of rules that track well ECB monetary policy decisions made in the period 1999–2010. Nevertheless, both identify episodes of significant deviations. Most interestingly the relevant periods include the Global Financial and Economic Crises.

In the NAWM deviations of actual policy rates (captured in the model by the Euro Interbank Offered Rate (EURIBOR)) from the levels implied by the rule were positive during 2008 and the beginning of 2009. In other words, policy rates were higher than implied by the rule. It can be seen from Fig. 4, which decomposes the contributions of various shocks to growth, that from 2008:2 to 2009:1 there is a negative contribution of deviations of monetary policy from benchmark to GDP growth. However, during this period, the three-month EURIBOR reflected a substantial risk premium, while the EONIA became substantially lower than the MRO rate during the period (see Fig. 3). Time-varying risk premia and financial intermediation spreads explain departures from the stylized transmission mechanism that obtains in models with a single interest rate. Those departures are particularly important in times of financial stress. Moreover, the monetary policy rule captures only monetary policy stance as far as it is reflected in interest rates. During this period the ECB introduced non-standard measures aiming at compensating for the malfunctioning of some aspects of the monetary transmission mechanism and the zero bound on nominal interest rates (see 'Monetary policy measures during the Global Financial Crisis', below). Orphanides (2010) comments on the same episode on the basis of the similar indications obtained from the simple first-difference rule.

The discussion makes clear that a variety of simple rules do help to convey the systematic character of monetary policy making. They do account *ex post* for a substantial part of monetary policy decisions taken. Nevertheless, it is important to avoid the mistake of thinking that they can substitute for the monetary policy strategy itself. As the above quote makes clear, the relevant considerations in a particular situation cannot be known in advance. No simple rule can come close to summarizing the full set of relevant considerations. The stability-oriented monetary policy strategy of the ECB strives to include the full set of relevant information filtered through a diversity of analytical perspectives.

Communication, Accountability and Transparency

Communication, accountability and transparency are today of crucial importance for central banking. This is the case for two basic reasons. First, in a democratic society independence goes together with accountability. In the euro area, the ECB has full independence in the pursuit of price stability. At the same time it is made accountable for the results achieved through a number of statutory

requirements (in article 284.3 of the Treaty on the Functioning of the European Union and article 15 of the Statute). These include the obligation publish an annual report on monetary policy and other activities of the system. The report is presented, by the President of the ECB, to the European Parliament and to the Economic and Financial Affairs Council (ECOFIN). The ECB must also prepare quarterly reports and weekly consolidated financial statements for the ESCB. Second, the transmission of monetary policy to prices and economic activity depends on private sector expectations. A central bank controls only very short-term interest rates. However, what matters most for the transmission of monetary policy impulses to the economy are longer-term interest rates that are determined by financial markets. Those longer-term rates and financial prices reflect market expectations of future short-term interest rates and premiums for uncertainty. In an uncertain and complex environment, monetary policy decisions do not necessarily convey the central bank's overall assessment of the current state of, and future prospects for, the economy. Hence monetary policy actions can only be properly understood within the broader context of the monetary policy strategy. Such requirement implies a constant effort of communication on how individual decisions and the monetary policy stance contribute to the achievement of the mandatory goals. By stabilizing market expectations good communication and transparency help to reduce uncertainty and volatility in financial markets. This reduces risk premia in real interest rates to the benefit of overall economic welfare.

The ECB went much beyond the statutory accountability requirements and communication constitutes a central element of ECB monetary policy-making. As an important example we have already emphasized above the ECB's decision to inform the public comprehensively before the start about its monetary policy strategy including a quantitative definition of price stability, based on a specific statistical indicator (HICP for the euro area).

Monetary policy decisions are communicated in real time. The key elements of the ECB's communication policy are the press conference held by the President and the Vice-President after the first Governing Council meeting of each month and the *Monthly Bulletin*. The introductory statement read out by the president provides a summary of the policy-relevant assessment of economic developments in line with the ECB's monetary policy strategy. This assessment is agreed by the Governing Council and contains the core message for communication with the markets, the media and the public. This assessment is typically reflected in numerous statements and speeches by Council members. The introductory statement is immediately followed by questions and answers. A transcript of the statement is published in all EU languages within a few hours. A week later the ECB publishes its *Monthly Bulletin*, which includes a thorough analysis of developments in the euro area and the global environment. It contains also all relevant statistical data.

The ECB also publishes economic projections regularly (see above). Detailed information on the models used in the projection exercises and in policy analysis is also released and made available through the ECB's web site.

An early and constant critique of the ECB's communication policy refers to the fact that voting records are not published (Buiter (1999); for a rejoinder see Issing (1999b), which also includes a survey of practices of major central banks). As a matter of fact, monetary policy decisions of the Governing Council are taken by consensus. This underlines the collective responsibility of the decision-making body. Any attempt to make individual policymakers personally accountable entails the risk that the public may attach more importance to individual opinions than to the relevant economic arguments. Particularly in a monetary union constituted of many countries, the behaviour of national central bank governors might be interpreted from a 'national' perspective (Issing 2005b).

Blinder et al. (2008) show that there is no consensus on best practices in central banking communication. Ehrmann and Fratzscher (2007) focused on the specific experience of the ECB, with press conferences as vehicles for explanation of monetary policy decisions. They looked at evidence from financial markets. They found

that press conferences provided substantial additional information to financial markets and with relatively low effects on volatility.

Monetary Policy Measures During the Global Financial Crisis

The global nature of the crisis made its first appearance on Thursday 9 August 2007. In the morning, traded volumes fell sharply in money markets, while interest rates suffered a sudden and significant increase to elevated levels, well above the ECB's minimum bid rate. In this context, the ECB was the first central bank to take action: it immediately provided liquidity through a fine-tuning operation. The ECB distinguishes the **monetary policy stance** from **monetary policy implementation**. The former, in normal circumstances, can be gauged on the basis of a money market interest rate. Monetary policy implementation, in contrast, is performed in the context of the operational framework, and is used, for example, to maintain orderly money market conditions and adequate provision of liquidity to banks.

On 3 July 2008, the ECB announced that it had decided to increase its key interest rates by 25 basis points. The decision was based on the ECB's assessment of the prospects for price developments and risks to price stability. The exchange of views between Petra Geraats and Frank Smets (Geraats 2010; Smets 2010) highlights the importance of private sector inflation expectations in this context.

However, the situation and prospects changed rapidly over the summer. Clear signals of a sharp economic slowdown in the USA and elsewhere became apparent. More dramatically, from September 2008, a perverse feedback spiral between economic and financial developments threatened to take hold. The failure of Lehman Brothers on 15 September 2008 became the emblematic event, marking the transition to the acute stage of the Global Crisis.

Again the ECB reacted rapidly and forcefully.

It is possible to summarize all measures taken by the ECB under five points:

1. Adjustment in key interest rates: interest rates were lowered 325 basis points from October 2008. For example, the minimum bid rate was lowered from 4.25% to 1%.
2. Liquidity support mechanisms; adjustments in the operational framework:
   a. Extended use of fine-tuning operations;
   b. Conduct of fixed rate tenders with full allotment;
   c. Expansion of the list of eligible collateral;
   d. Temporary narrowing of the interest rate corridor;
   e. Lengthening of maturity for Long-Term Refinancing Operations.
3. Acquisition of selected assets: purchase of euro-denominated covered bonds (under the Covered Bond Purchase Programme (CBPP)) and interventions in euro area public and private debt markets (under the Securities Markets Programme (SMP)).
4. Joint action with other central banks on October 2008 to announce a reduction in interest rates.
5. Cooperation with other central banks in the management of liquidity in foreign currencies.

This episode suggests that the Eurosystem made use of the ample flexibility afforded by the operational framework to meet the special challenges associated with the crisis.

Challenges Going Forward

In this article we have argued that the conduct of monetary policy has been effective in the first years of the euro area. Nevertheless the ongoing crisis has brought into sharp focus a number of fundamental questions that will mark developments going forward. These include (for a longer list see Gaspar (2010)):

- Will the ECB manage successfully the exit from its current exceptional stance and continue its impressive record in maintaining price stability over the medium term?
- Will the framework for financial supervision and regulation prove effective? Will the new

European Financial Stability Risk Board and the European System of Financial Supervision work well? How will the ECB, in particular, and central banks, in general, adapt to the new systemic risk management framework?
- Will the single currency continue to be an important driver of deeper integration in the single market?
- Can rules and procedures, aiming at fiscal discipline in the euro area, effectively mitigate the deficit bias in government finance and ensure sound public finances in view of the ongoing demographic transition and of the need to provide for fiscal space?
- How to protect financial stability, of the euro area as a whole, in the face of turmoil in sovereign debt markets?

In more general terms, the challenge derives from the combination of a single market and a single currency with national responsibilities in the areas of economic policy and financial stability.

See Also

- ▶ Central Bank Communication
- ▶ Central Bank Independence
- ▶ European Central Bank
- ▶ Taylor Rules

Bibliography

Alessi, L., and C. Detken. 2009. 'Real time' early warning indicators for costly asset price boom/bust cyces: A role for global liquidity. ECB Working Paper No. 1039.

Amisano, L., and G. Fagan. 2010. Money growth and inflation: a regime switching approach. ECB Working Paper 1207. http://www.ecb.int/pub/pdf/scpwps/ecbwp1207.pdf

Amisano, G., A. Beyer, and M. Lenza. 2010. Enhancing monetary analysis. *ECB Research Bulletin 11*. http://www.ecb.int/pub/pdf/other/researchbulletin11en.pdf

Barro, R.J., and D. Gordon. 1983. A positive theory of monetary policy. *Journal of Political Economy* 98: S103–S125.

Beck, G., and V. Wieland. 2008. Central bank misperceptions and the role of money in interest rate rules. *Journal of Monetary Economics* 55: S1–S17.

Beyer, A. 2009. A stable model for euro area money demand: Revisiting the role of wealth. ECB Working Paper 1111. http://www.ecb.int/pub/pdf/scpwps/ecbwp1111.pdf

Bindseil, U. 2004. *Monetary policy implementation*. Oxford: Oxford 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 theory and evidence. ECB Working Paper Series 898. http://www.ecb.int/pub/pdf/scpwps/ecbwp898.pdf

Buiter, W.H. 1999. Alice in Euroland. *Journal of Common Market Studies* 37(2): 181–209.

Buiter, W. H. 2009. Central banks and financial crises. Paper presented at Maintaining Stability in a Changing Financial System, Jackson Hole, 21–23 August 2008. http://www.kansascityfed.org/publicat/sympos/2008/Buiter.03.12.09.pdf

Buti, M., S. Deroose, V. Gaspar, and J. Nogueira Martins. 2010. *The euro: The first decade*. Cambridge: Cambridge University Press.

Christoffel, K., G. Coenen, and A. Warne. 2008. A new area wide model for the euro area: A micro-founded open-economy model for forecasting and policy analysis. *ECB Working Paper 944*. http://www.ecb.int/pub/pdf/scpwps/ecbwp944.pdf?35c290946f2f220119ec24d3a5394ebd

Clarida, R., J. Galí, and M. Gertler. 1999. The science of monetary policy: A new Keynesian perspective. *Journal of Economic Literature* 37: 1661–1707.

Ehrmann, M., and M. Fratzscher. 2007. Explaining monetary policy in press conferences. *ECB Working Paper 767*. http://www.ecb.int/pub/pdf/scpwps/ecbwp767.pdf

EMI. 1997a. *The single monetary policy in stage three: Specification of the operational framework*. http://www.ecb.int/pub/pdf/othemi/pub_05en.pdf.

EMI. 1997b. *The single monetary policy in stage three: Elements of the monetary policy strategy of the ESCB*. http://www.ecb.int/pub/pdf/other/singlemonetarypolicystagethreeelementsmonetarypolicystrategyescb1997en.pdf

EMI. 1997c. *Annual report*. http://www.ecb.int/pub/pdf/annrep/ar1997en.pdf.

EMI. 1998. *The single monetary policy in stage three: General documentation on the ESCB monetary policy instruments and procedures*. http://www.ecb.int/pub/pdf/other/gendoc98en.pdf

European Central Bank. 1998. Press release: A stability-oriented monetary policy strategy for the ESCB, 13 October. http://www.ecb.int/press/pr/date/1998/html/pr981013_1.en.html

European Central Bank. 1999a. The euro area at the start of stage three. *Monthly Bulletin*, January. http://www.ecb.int/pub/pdf/mobu/mb199901en.pdf

European Central Bank. 1999b. The stability-oriented monetary policy strategy of the Eurosystem. *ECB Monthly Bulletin*, January. http://www.ecb.int/pub/pdf/mobu/mb199901en.pdf

European Central Bank. 2000. The two pillars of the ECB monetary policy strategy. *ECB Monthly Bulletin*,

November. Available at http://www.ecb.int/pub/pdf/mobu/mb200011en.pdf
European Central Bank. 2001. *The monetary policy of the ECB*. Frankfurt-am-Main: ECB.
European Central Bank. 2003a. Press release: The ECB's monetary policy strategy. http://www.ecb.int/press/pr/date/2003/html/pr030508_2.en.html
European Central Bank. 2003b. The outcome of the ECB's evaluation of its monetary policy strategy. *ECB Monthly Bulletin*, June. http://www.ecb.int/pub/pdf/mobu/mb200306en.pdf
European Central Bank. 2008. *The implementation of monetary policy in the euro area: General documentation on the eurosystem's policy instruments and procedures*. Frankfurt-am-Main: ECB. http://www.ecb.europa.eu/pub/pdf/other/gendoc2008en.pdf
European Central Bank. 2010. Enhancing monetary analysis. *ECB Monthly Bulletin*, November. http://www.ecb.int/pub/pdf/mobu/mb201011en.pdf
Fagan, G., J. Henry, and R. Mestre. 2001. An area-wide model (AWM) for the euro area. ECB Working Paper 42. http://www.ecb.int/pub/pdf/scpwps/ecbwp042.pdf
Friedman, M. 1969. The optimum quantity of money. In *The optimum quantity of money and other essays*. Chicago: Aldine Publishing Company.
Galí, J. 2008. *Monetary policy, inflation and the business cycle: An introduction to the new Keynesian framework*. Princeton: Princeton University Press.
Gaspar, V. 2010. *The euro second decade: Success continues!* http://www.dallasfed.org/institute/events/2010/10euro_gaspar.pdf
Geraats, P. 2010. ECB credibility and transparency. In *The euro: The first decade*, ed. M. Buti, S. Deroose, V. Gaspar, and J. Nogueira Martins. Cambridge: Cambridge University Press.
Issing, O. 1999a. The eurosystem: Transparent and accountable or "Willem in Euroland". *Journal of Common Market Studies*, September, 503–519.
Issing, O. 1999b. The monetary policy of the ECB in a world of uncertainty. In: *Monetary policy making under uncertainty*, ed. I. Angeloni. http://www.ecb.int/press/key/date/1999/html/sp991203_2.en.html
Issing, O. 2000. Monetary policy in a new environment. Speech delivered at the Bundesbank-BIS conference on recent developments in financial systems and their challenges for economic policy: An European perspective, Frankfurt-am-Main. http://www.ecb.int/press/key/date/2000/html/sp000929.en.html
Issing, O. 2005a. The monetary pillar of the ECB. Presented at The ECB and its Watchers VII, Frankfurt, 3 June. http://www.ecb.int/press/key/date/2005/html/sp050603.en.html
Issing, O. 2005b. Communication, transparency, accountability: Monetary policy in the twenty-first century. *Federal Reserve Bank of St. Louis, Review*, March/April, 65–83.
Issing, O. 2008. *The birth of the euro*. Cambridge: Cambridge University Press.
Issing, O., V. Gaspar, I. Angeloni, and O. Tristani. 2001. *Monetary policy in the euro area: Strategy and decision-making at the European Central Bank*. Cambridge: Cambridge University Press.
Issing, O. ed. in cooperation with I. Angeloni, V. Gaspar, H.-J. Klöckers, K. Masuch, S. Nicolleti-Altimari, M. Rostagno, and F. Smets 2003. *Background studies for the ECB's evaluation of its monetary policy strategy*. Frankfurt-am-Main: ECB. http://www.ecb.int/pub/pdf/other/monetarypolicystrategyreview_backgrounden.pdf
Issing, O., V. Gaspar, O. Tristani, and D. Vestin. 2005. *Imperfect knowledge and monetary policy, the stone lectures in economics*. Cambridge: Cambridge University Press.
Jonung, L., and E. Drea. 2010. It can't happen, it's a bad idea, it won't last. US economists on the EMU and the euro; 1989–2002. *Econ Journal Watch* 7(1): 4–52.
Kydland, F.E., and E.C. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85(3): 473–491.
Lucas, R.E. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* I: 19–46.
McCallum, B. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester Series on Public Policy* 29: 173–204.
McCallum, B. 1999. Issues in the design of monetary policy rules. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford, vol. 1c. Amsterdam: North-Holland.
Minsky, H. 1975. *John Maynard Keynes*. Columbia: Columbia University Press.
Neumann, M. 2010. Some observations on the ECB monetary policy. In *The euro: The first decade*, ed. M. Buti, S. Deroose, V. Gaspar, and J. Nogueira Martins. Cambridge: Cambridge University Press.
Official Journal of the European Union. 2008. *Consolidated versions of the treaty on European Union and the treaty on the functioning of the European Union*, 9 May.
Orphanides, A. 2003. Historical monetary analysis and the Taylor rule. *Journal of Monetary Economics* 50(5): 983–1022.
Orphanides, A. 2006. Contribution to the panel ECB Watch: Review of the ECB monetary policy strategy and alternative approaches. CFS Research Conference: the ECB and its Watchers VIII, Frankfurt, 5 May. https://www.ifk-cfs.de/index.php?id=697
Orphanides, A. 2010. Monetary policy lessons from the crisis. The great financial crisis: Lessons for financial stability and monetary policy: Colloquium in honour of Lucas D. Papademos, Frankfurt, 20–21 May. http://www.ecb.europa.eu/pub/pdf/other/greatfinancialcrisisecbcolloquiumpapademos201203en.pdf
Papademos, L., and J. Stark. eds. 2010. *Enhancing monetary analysis*. Frankfurt-am-Main: ECB. http://www.ecb.int/pub/pdf/other/enhancingmonetaryanalysis2010en.pdf
Romer, D. 2006. *Advanced macroeconomics*. 3rd ed. New York: McGraw-Hill.

Smets, F. 2010. Comment 3: Comment on Chapters 6 and 7. In *The euro: The first decade*, ed. M. Buti, S. Deroose, V. Gaspar, and J. Nogueira Martins. Cambridge: Cambridge University Press.
Smets, F., and R. Wouters. 2002. An estimated stochastic dynamic general equilibrium model of the euro area1. ECB Working Paper 171. http://www.ecb.int/ pub/pdf/scpwps/ecbwp171.pdf?88101db0141ddff8c6d6d0ed3cb823b5
Stark, J., and L. Papademos. eds. 2010. *Enhancing monetary analysis*. Frankfurt-am-Main: ECB. http://www.ecb.int/pub/pdf/other/enhancingmonetaryanalysis2010en.pdf
Taylor, J. 1993. Discretion versus policy rules in practice. Carnegie-Rochester Conference Series on Public Policy, 39, December, 195–214.
Walsh, C. 2010. *Monetary theory and policy*. 3rd ed. Cambridge: MIT Press.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

European Cohesion Policy

Willem Molle

Abstract

One of the main objectives of the EU is cohesion, namely a decrease in the disparity of wealth between its constituent parts. A considerable part of the EU budget is earmarked for this policy. The policy should also support the EU 2020 strategy, which focuses on smart, inclusive and sustainable growth. Over the past decades the EU has been able to realize to some extent its ambitions. Effectiveness can be further stepped up by making improvements in the delivery system.

Keywords

Convergence; Competitiveness; Coordination; Disparity; Effectiveness of intervention; Regulation; Principles; Structural Funds

JEL Classifications

R1; R53; R58

Why EU Involvement?

The European Union has to contend with enormous disparities in wealth between the Member States. Successive enlargements have considerably increased this inequality due to the accession of low income new Member States. These disparities lead to social and political problems that endanger the internal cohesion of the EU. Moreover, the very dynamics of the integrated economy of the EU may lead to further agglomeration and hence increase cohesion problems. Hence a policy is needed to change the situation and bend the autonomous development processes in such a way that they lead to less disparity and more cohesion. In principle the national governments of the Member States are the first in line to cope with these problems.

However, there are two sound reasons (related to the subsidiarity principle) why the EU should also step in. One important consideration is the economies of scale applied to finances: the EU can mobilize and provide greater funds under far better conditions than poor Member States. Moreover, it can offer long-term predictability about the availability of resources to all beneficiaries. This means that investors will be more inclined to invest and growth is therefore likely to be enhanced. The second important factor is regulation: the EU sets rules to limit internal competition between Member States which offer state aid. Moreover, the EU determines the architecture and operation of the delivery system while leaving Member States to oversee the application of the eligibility criteria and the selection of projects within the EU priorities. (For an elaborate systematic analysis of all stages of the policy and for detailed references to the relevant literature, see Molle (2007)).

Definition and Measurement of Cohesion

Cohesion is a concept that has been introduced into EU policy without a precise definition. Over time a practical definition has been developed. Cohesion has increasingly become understood as the degree to which disparities in social and

economic welfare between the different regions or groups within the European Union are politically and socially tolerable. Whether cohesion is achieved is largely a political question.

Cohesion is measured by the change in disparity from one period to another. A decrease in disparity (convergence) means improved cohesion, whereas an increase in disparity (divergence) means less cohesion. In general one uses simple indicators to measure disparity; the most common one is regional Gross Domestic Product per head. Another common indicator is (un-) employment. Moreover, a series of other indicators are used, such as the risk of poverty, health, and access to broadband Internet.

Objectives of the Policy

The EU has set a number of objectives for its cohesion policy. The fundamentals have remained fairly constant over time, although the specifics have constantly been adapted to new challenges (Begg 2010). The main ones are:

1. To improve cohesion (that is convergence of wealth levels) on three dimensions:
   - economic – i.e. the conditions for economic growth such as innovation
   - social – i.e. employment and social exclusion (e.g. poverty)
   - territorial – i.e. specific types of regions (such as urban), elements directly related to spatial planning (e.g. infrastructure) and the environment (Duehr et al. 2010).
2. To contribute to other EU objectives, for instance:
   - facilitating major advances in economic integration, such as enlargement or the passing on to higher stages of integration (e.g. Economic and Monetary Union)
   - contributing to major policy targets such as the increase in competitiveness, the decrease of social exclusion or the stimulation of environmental sustainability. (The latter has notably come to the fore with the so called Lisbon strategy, launched in 2000, and the new Europe 2020 strategy (EC 2010a).)

The objective discussed under 1 is generally referred to as the *convergence* objective. The regions that fall under this objective have a GDP/P level below 75% of the EU average. They are the main beneficiaries of the policy. The EU has decided to let all other regions also benefit from cohesion policy; the main objectives here are the improvement of competitiveness and social inclusion (these regions are therefore generally called *competitiveness* regions). (Up till the present programming period the former type regions were formally called objective 1 regions and the second type objective 2.)

Instruments

The main instruments by which the cohesion policy is put into effect are:

1. The provision of financial means (see the following section). The EU does this by allocating funds to the disadvantaged regions to improve their economic structure and to social groups to improve their employability and to avoid their social exclusion. Both should lead to increases in competitiveness.
2. The setting of rules and the coordination of actions (see 'Regulation and coordination', below). As cohesion is a matter of shared responsibilities between the Union and national authorities, such coordination is vital for effectiveness. This applies equally to national cohesion policies and to other EU and national policies, such as environment.

Financial Support

The main instrument of cohesion policy is financial support paid by the EU budget. As a first step, a share of the budget (some 40%) is earmarked for cohesion. In the next step, funds are allocated to the different Funds (see next paragraph). In the third step the former are allocated to the various cohesion objectives. Convergence regions (defined as those with an average wealth level of less than 75% of the EU mean) get the lion's share

(three quarters). They are mostly located in the new Member States and in the Mediterranean South. The rest is for regions anywhere else in the EU to improve their competitiveness and for territorial cooperation. This step also defines the allocation of resources over countries. In the fourth step a selection is made of the programmes and projects that are eligible for support.

Spending on cohesion operates mainly through two types of fund:

- The European Regional Development Fund (ERDF) and the European Social Fund (ESF). These funds are called 'structural' because they support measures that aim at the improvement of the structural aspects of the regional economy. There is a certain specialization between the two funds: the ERDF concentrates on economic cohesion and finances mainly infrastructure and innovation; the ESF concentrates on social cohesion and finances mainly training and education.
- The Cohesion Fund (CF). Beneficiaries are the member countries with below EU average (actually 90%) GDP per head figures, with a programme of economic convergence to EMU conditions. The Cohesion Fund finances environmental and transport projects in a framework that is different from the Structural Funds; it delivers national, not regional funding and the programming is simplified compared to the ERDF and the ESF.

Note that these funds provide only financial support to projects and programmes. The principle of additionality prescribes that major contributions to the financing have to be made by the national and or regional governments too. The lower the wealth level of the region, the higher, in percentage terms, is the EU contribution.

Main Actors

The EU is not the only body responsible for cohesion. Member States also play an important role. Moreover, according to the partnership principle the policy involves local governments and representatives of the third and private sector. This multi-level governance of the EU cohesion policy is meant to increase participation and coordination and hence consistency and effectiveness. The competences of the three main actors, and hence the power balance between them, changes in the course of the policy cycle (see Table 1).

A whole administrative and institutional system has been set up to deliver the policy. This consists first of Management Authorities, which are responsible for the programming and execution of the various projects, such as the building of roads and the training of people. They also guide the work of the Monitoring Committees, which are the responsibility of the Member States. They are the highest decision-making bodies of each Operational Programme. They are charged with surveillance of the progress of the work; reporting to the European Commission and making proposals for adjustments. The financial disbursements are done by so-called Certifying Authorities, while control on spending is done by Auditing Authorities.

European Cohesion Policy, Table 1 Changing roles of the major actors during the policy cycle (adapted from Molle (2007), p. 127)

Stage in the policy cycle	European Commission	National governments (Council)	Regional authorities
Basic design	strong	dominant	weak
Financial packages Definition of objectives and of eligibility criteria	modest	dominant	weak
Institutional framework and delivery system	strong	strong	modest
Implementation	weak	variable	strong
Evaluation	strong	variable	variable

Regulation and Coordination

The EU cohesion policy also uses the instrument of regulation to set uniform rules that govern the use of financial instruments (the Structural Funds; see Instruments, above) and to set the framework for the coordination among partners (see previous section) to realize mutual consistency of objectives, priorities and concrete projects.

Moreover, the instrument of regulation is used where one can assume that the coordination instrument is insufficiently effective at controlling certain negative effects of independent national and regional policy making. This applies notably to state aid. The main principle of EU competition policy sets a complete ban on state aid. However, this ban can be lifted in the case of regional structural weaknesses. The degree to which this is possible has been made dependent on the level of development of the Member States: restricted in the case of rich Member States and somewhat more lenient in that of poor Member States. This rule serves the purpose of effectiveness; subsidy wars would always be won by the richer Member States, as they have the resources to outbid the poorer Member States, which would render ineffective the EU support given to the latter.

Consistency with Other EU Policies

The EU pursues a large number of policies, some of which can have a significant influence on the distribution of economic activity, which in turn can have an adverse effect on cohesion. To avoid such problems the design and implementation of major EU policies have to be coordinated. This is notably the case for policies with a strong spatial or territorial dimension, such as transport and environment, but also for policies dealing with research and the information society. Indeed, innovation being one of the main determinants of competitiveness and hence growth, EU-wide policies and national and regional policies in this domain have to be dovetailed to achieve maximum effect (Molle 2009). It is to be noted here that the policies of the EU that are fundamental for the functioning of the internal market (such as freedom of movement) and Monetary Union (such as the Stability and Growth Pact) are spatially blind; their potential negative effects on cohesion have to be compensated for by an increased focus on cohesion policy in the affected areas.

Integrated and Territorial Approach

The coordination of sectoral policies in a multilevel government framework is particularly difficult (Meijers and Staed 2004). Indeed, the search for compromises at the EU level on cross-cutting issues tends to lead to vertical inconsistencies at the national and regional levels, and other compromises between sectoral policies may have been worked out. Vertical policy integration by sector may complicate the solution of intersectoral conflicts on the EU, national or local levels. As part of a solution to these simultaneous problems the EU has decided to make Impact Assessments of its policy proposals and put some accent on territorial impact assessments. Moreover, it has opted for integration of its policies at the regional and local levels, as at this level things move from the abstract into the concrete; indeed general objectives have to be translated into concrete projects on the basis of priorities that are selected with a view to their local potential (Barca 2009; EC 2010b).

Evaluation of Effectiveness

The performance of the EU cohesion policy (that is realization of its objectives) is difficult to evaluate due to methodological and practical insufficiencies. Over the past decades much effort has been made both by academics, consultants and policy makers to improve the situation (Basle 2006; Mairate 2006; Martin and Tyler 2006). However, these have not resulted in a consensus as to its effectiveness. On the contrary one sees two almost opposite views.

The *dominant view is positive*. Those who hold this view have found that in the past the EU cohesion policy has been effective in decreasing the wealth gap between its member countries and regions (Bornschier et al. 2004; Ederveen et al. 2003; Martin and Saenz 2003; Tselios 2009). Moreover, due to the interrelations between different member countries the net payers have benefited also because their industries have been the main suppliers of investment goods to the projects executed under the cohesion policy. Next, they indicate the contribution to a series of side objectives, such as the environment (ENEA 2006). Finally they stress that the EU has increased effectiveness by setting up a better intervention and delivery system than the national ones (Lion et al. 2004; Gualini 2004), geared to local needs and capacities (Leonardi 2005). These results are embraced by the Commission (EC 2010c).

There are also (*highly*) *critical views*. Some find that cohesion policy is not appropriate: a good set of other policies producing the conditions for healthy growth would also lead to convergence of wealth levels. Others find that cohesion policy is not effective; they observe that during the years that the EU cohesion policy was still very limited, disparities decreased, while they have increased since the huge increases in EU cohesion spending (Boldrin and Canova 2003; Dall'Erba and Gallo 2008). Others find that the policy is inefficient; the actual management of the policy takes up too large a share of resources.

Conclusions and Future Developments

Although there is no clear-cut conclusion from empirical research as to the performance of the policy, there is much evidence to support the view that over the past decades the EU cohesion policy has been conducive to the attainment of major EU policy objectives. This applies in particular to the convergence objective and some side objectives, such as the internal market and the EMU. All is not rosy however; in the course of time certain weak points have become apparent that need adaptation. We may mention the dependency on aid of the convergence countries, the welfare loss incurred in transferring money from rich member states back to their regions via the 'competitiveness' route, the rigidity of pre-fixed quota and priorities, the high delivery cost for atomized projects notably of the European Social Fund, etc.

To remedy such problems many proposals have been made for a renewed structure of the EU cohesion policy. Some of them are not new; indeed, the inadequacies of the basic architecture of the policy have been apparent for some time (Bachtler and Mendez 2007). The most far-reaching proposals include the abolition of the 'competitiveness' objective or at least the redirection of these funds to the respective budget headings (implied in the subsidiarity tests as given in, for example, Begg (2008) and ECORYS (2008)). Another proposal is the abolition of the quota system of the competitiveness objective and its replacement by a system of competitive bidding for EU priority projects (Ederveen et al. 2003; Tarschys 2003). Finally, proposals have been made to remedy the potential negative effects on aid dependency by much stricter conditionality in matters of administrative capacity (Molle 2011).

Much like in the previous round of reforms of the cohesion policy the European Commission has maintained as an essential feature of the policy that all regions may benefit. The reasons are of a political nature. First, the regions of rich Member States are not inclined to give up access to EU funds. Second, the Commission is not inclined to give up its influence in all regions of the EU, the less so because it considers that the EU 2020 strategy justifies continuation of the competitiveness objective – this notwithstanding the general view that the use of the cohesion policy instruments for such purposes is a rather distorted way of matching objectives with instruments.

So, the proposals of the Commission for adaptation of the policy in the period 2013–2020 focus on delivery aspects such as the strategic programming (see also Barca 2009). They follow certain of the proposals mentioned earlier by critics, such as thematic concentration, conditionality of

support, evaluation of impact, the use of new financial instruments (not only grants but also loans), streamlining of the financial management and control systems and finally strengthening of the institutional capacity of the recipients (EC 2010d).

See Also

- European Labour Markets
- European Monetary Union
- Regional Development, Geography of
- Regional Distribution of Economic Activity
- Regional Economics
- Spatial Economics

Bibliography

Bachtler, J., and C. Mendez. 2007. Who governs EU cohesion policy? Deconstructing reforms of the Structural Funds. *Journal of Common Market Studies* 45(3): 535–564.

Barca, F. 2009. *An agenda for a reformed cohesion policy: A place based approach to meeting European Union challenges and expectations*. Brussels (available from the Commission website; DG Regio).

Basle, M. 2006. Strengths and weaknesses of European Union policy evaluation methods; ex post evaluation of Objective 2, 1994–1999. *Regional Studies* 40(2): 225–235.

Begg, I. 2008. Subsidiarity in regional policy. In *Subsidiarity and economic reform in Europe*, ed. G. Gelauff, I. Grilo, and A. Lejour, 291–310. Berlin: Springer.

Begg, I. 2010. Cohesion or confusion; a policy searching for objectives. *Journal of European Integration* 32(1): 77–96.

Boldrin, M., and F. Canova. 2003. Regional policies and EU enlargement. *CEPR discussion paper series*, no. 3744.

Bornschier, V., M. Herkenrath, and P. Ziltener. 2004. Political and economic logic of Western European integration: A study of convergence comparing member and non member states 1980–1998. *European Societies* 6(1): 71–96.

Dall'Erba, S., and J. Le Gallo. 2008. Regional convergence and the impact of European Structural Funds over 1989–1999: A spatial econometric analysis. *Papers in Regional Science* 87(2): 219–244.

Duehr, S., C. Colomb, and V. Nadin. 2010. *European spatial planning and territorial cooperation*. London: Routledge.

EC. 2010a. *Europe 2020: A strategy for smart, sustainable and inclusive growth*. COM 2020, 3 March.

EC. 2010b. Regional policy, an integrated approach. *Panorama Inforegio,* 34.

EC. 2010c. *Investing in Europe's future: Fifth report on economic, social and territorial cohesion*. Brussels: EC.

EC. 2010d. *Conclusions of the fifth report on economic, social and territorial cohesion: The future of cohesion policy,* (SEC 2010) 1348 final. Brussels: EC.

ECORYS/CPB/Ifo. 2008. *A study on EU spending*, Final report. Rotterdam (available from the website of the EU Commission).

Ederveen, S., J. Gorter, R. de Mooy, and R. Nahuis. 2003. Funds and games; the economics of European cohesion policy. *ENEPRI Occasional paper no. 3*. .http://www.enepri.org/.

ENEA. 2006. *The contribution of structural and cohesion funds to a better environment* (available via the EC website).

Gualini, E. 2004. *Multi-level governance and institutional change: The Europeanization of regional policy in Italy*. Aldershot: Ashgate.

Leonardi, R. 2005. *Cohesion policy in the European Union: The building of Europe*. Basingstoke: Palgrave.

Lion, C., P. Martini, and S. Volpi. 2004. Evaluation of European Social Fund Programmes in a new framework of multinational governance; the Italian experience. *Regional Studies* 38(2): 207–212.

Mairate, A. 2006. The 'added value' of European cohesion policy. *Regional Studies* 40(2): 167–178.

Martin, C., and I. Saenz. 2003. Real convergence and European Integration; the experience of the less developed EU members. *Empirica* 30(3): 205–236.

Martin, R., and P. Tyler. 2006. Evaluating the impact of the Structural Funds on Objective 1 regions, an exploratory discussion. *Regional Studies* 40(2): 201–210.

Meijers, E., and D. Staed. 2004. Policy integration; what does it mean and how can it be achieved?; A multidisciplinary review. *Berlin Conference on the Human Dimensions of Global Environmental Change, Greening of Policies, Interlinkages and Policy Integration*. http://www.un.org/esa/desa/papers/2009/wp73_2009.pdf.

Molle, W. 2007. *European cohesion policy.* London: Routledge.

Molle, W. 2009. European innovation policy; increased effectiveness through coordination with cohesion policy. In *Enhancing the effectiveness of innovation in Europe; new roles for key players*, ed. W. Molle and J. Djarova, 167–200. Cheltenham: Edward Elgar.

Molle, W. 2011. *Europe Post 2013? Limit Ambitions; Step up Capacity!* (paper for the April 2011 Bled conference of the Regional Studies Association, available from their website).

Tarschys, D. 2003. Reinventing cohesion: The future of European Structural Policy, *Report no 1.* Stockholm: Sieps.

Tselios, V. 2009. Growth and convergence in income per capita and income inequality in the regions of the EU. *Spatial Economic Analysis* 4(3): 344–371.

European Employment Policy

Robert Strauss
European Commission, Brussels, Belgium

Abstract

The European Employment Strategy (EES) was included in the 1997 Treaty of Amsterdam to accompany the newly completed Single Market and the advanced preparations for the European Monetary Union. It constitutes the core of European employment policy. The entry examines how and why the EES came into being and the employment policy challenges it has focused on since its inception until today. It looks at some of the labour economics underpinning it and at what the EES but also other employment policies have been able to achieve in the EU, especially since the economic and financial crisis of 2008.

Definition

European employment policy is the set of actions taken at the level of the European Union to enhance the quantity and quality of jobs in all the Member States.

Introduction

Employment has been a key concern and indeed an objective of the European integration process since the Treaty of Rome establishing the European Economic Community was signed in 1957. In its preamble its signatories declared that they are "resolved to ensure the economic and social progress of their countries. . .. affirming as the essential objective of their efforts the constant improvement of living and working conditions. . .." The European Social Fund providing financial support for improving workers' mobility and employment opportunities in the common market was also established (EEC 1957). But it was only at the Paris Summit of 1972 that it was decided to use the European Social Fund to support "the carrying out a co-ordinated policy for employment and vocational training, . . . improving working conditions and conditions of life, . . . closely involving workers in the progress of firms" (European Communities 1972). A European employment policy was emerging containing three pillars: a legislated rights pillar comprised of European employment legislation; "law via collective agreement" based on social partners' agreements; and thirdly, coordination of national employment policies (Rhodes 2005). Employment policy coordination became the basis for the European Employment Strategy, provided for in the Amsterdam Treaty (TEU 1997).

Included in the revised Treaty, to accompany the newly completed Single Market and the advanced preparations for the European Monetary Union (EMU), the European Employment Strategy (EES) constituted the crux of what is understood to be a, or the, European employment policy. It is buttressed by the legislative side, the first two pillars mentioned above, and the financial support provided by the European Social Fund. This entry will focus on how and why the EES came into being and the employment policy challenges it has focused on since its inception. In doing so, it will attempt to discuss some of the labour market economics underpinnings of the evolution of the EES. It will also look briefly at what it and other employment policies have or have not been able to achieve in the EU, especially since the economic and financial crisis of 2008.

The European Employment Strategy

The Delors White Paper and the Essen Process

The content of the Amsterdam Treaty establishing the European Employment Strategy did not come out of the blue. It was the result of the Delors White Paper on Growth, Competitiveness and Employment (Delors 1993) and the Essen European Summit of December 1994. And these two milestones had been preceded by several actions in the 1970s and 1980s with both legislative and Social Dialogue aspects to the fore

(Goetschy 1999). The Delors White Paper sought to offset the potentially deflationary convergence criteria of maximum shares of GDP of 60% for public debt and 3% for the fiscal deficit for the EMU enshrined in the Maastricht Treaty. In the employment field it pushed for greater use of active labour market policies (ALMPs) to complement increased labour market flexibility. ALMPs could be financed (unlike passive measures including unemployment benefits) from the Social Fund and the academic world was increasingly underlining their benefits (see e.g., Layard et al. 1991).

In Europe the employment situation had been deteriorating sharply. Between 1990 and 1994 the soon-to-become EU 15 saw six million net jobs lost. The unemployment rate rose from a cyclical low of 7.7–11.1%. Unemployment was at the top of the political agenda in many countries. The Essen summit of December 1994 certainly did not only focus on labour market issues, the accession process for Central and Eastern European countries was the key topic, but it did lay the groundwork for the EES. It agreed to set up a multilateral employment monitoring procedure closely modelled as the economic monitoring procedure contained in the Maastricht Treaty. Member States were recommended to take measures at national level in five areas:

- Improving employment opportunities by promoting investment in vocational training (especially for the young) and encouraging lifelong learning
- Increasing the employment intensity of growth, particularly through a more flexible organisation of work and working time, wage restraint, job creation in local environmental and social services
- Reducing non-wage labour costs to encourage employers to hire low-skilled workers
- Developing active labour market policies through the reform of employment services, encouraging occupational and geographical labour mobility and developing incentives for the unemployed to return to work
- Targeting measures to help groups particularly affected by long-term unemployment

Member States were urged to translate these recommendations into a long-term programme in the light of their specific economic and social circumstances and were required to submit an annual progress report. The Commission, in conjunction with the Economics and Financial Affairs Council (ECOFIN) and the Labour and Social Affairs Council, was to synthesise these national reports into an annual assessment submitted to the December European Council. On this basis, the European summit would review the employment guidelines, issue further recommendations to Member States and decide new initiatives at Community level.

"This procedure was intended to have a threefold effect. First, the annual report would help improve the efficiency of national employment policies by exposing these to public examination and facilitating explicit comparison of the performance of each Member State. Second, the prescribed cooperation between ECOFIN and the Social Affairs Council in drafting the annual report might facilitate greater integration of economic and employment policy, [essentially at a European level]. Third, it was hoped that multilateral employment monitoring would encourage greater convergence of employment policies in the Member States along the lines of the Essen recommendations" (Goetschy 1999, pp. 121–122).

The Amsterdam Treaty

The Amsterdam Treaty revising that of Maastricht made "full employment" an explicit priority of the EU and "a question of common concern." There was a realisation that the good or bad employment situation of one Member State affected or had spill-overs on another. The Community acquired new powers to develop "a coordinated strategy for employment" which should in particular promote "a skilled, trained and adaptable workforce and labour markets responsive to economic change."

The employment chapter of the Treaty covered: first, the integration of employment in the formulation and implementation of other Community policies; second, the establishment of mechanisms for coordinating employment policies at Community level. These mechanisms

reflected practices already in operation as part of the Essen monitoring procedure, but also borrow extensively from the economic policy coordination model setup by entry 103 of the Treaty of Maastricht. The main difference, and it is significant, is that recommendations issued on employment matters lack any binding effect.

In four significant respects, however, the Amsterdam Treaty involved an advance on the Essen process. First, the "annual guidelines for employment" were established as the driving force and the key component of coordination. Second, the Council carries out an annual examination of measures taken by Member States to implement the guidelines. The Council's evaluation is based on the annual report that each Member State must submit to the Council and the Commission, and on the opinion of the Employment Committee. The requirement to submit an annual programme was thus "hard" law and ultimately if not done could lead to fines being levied by the European Court. The quality of such programmes was not however specified. If necessary, in the light of this examination, the Council "acting by a qualified majority on a recommendation from the Commission, may, if it considers this to be appropriate, make recommendations to Member States." Such recommendations to individual states deemed not to have followed the guidelines would have no legally enforceable effect but are symbolically powerful. This implied a strengthening of influence at Community level.

Third, the Treaty establishes an Employment Committee with advisory status, formally ratifying an initiative taken by the Council in December 1996. The Member States and the Commission each appoint two members of the Committee. It has a dual purpose: to monitor the employment situation and employment policies in the Member States and the EU, and to formulate opinions (at the request of the Commission, the Council or on its own initiative) and to prepare the Council's work. In fulfilling its mandate it is required to consult the social partners.

Finally entry 5 of the Treaty allows the Council to adopt, by a qualified majority and after consulting the Economic and Social Committee and the Committee of the Regions, "incentive measures designed to encourage cooperation between Member States and to support their action in the field of employment." Such measures can involve the dissemination of best practice, the evaluation of experiences and the launch of pilot projects. The provisions of the Amsterdam Treaty came into force in May 1999, cementing practices known in Brussels circles as the Luxembourg process. Furthermore, the employment and labour ministers at the OECD in the 1997 ministerial meeting had endorsed "the need to shift public spending on labour market policies from passive to active measures" (OECD 1997).

The Luxembourg Process, the European Employment Strategy and the Lisbon Strategy

In November 1997, as unemployment remained high, an extraordinary ministerial Jobs Summit was held in Luxembourg. It launched the European Employment Strategy (EES) as set out in the Treaty, complemented by features which made it into the first open method of coordination (OMC) – the Luxembourg process. This process was the basis or model for similar soft law instruments in other social areas such as social protection, social inclusion and certain aspects of education and training. The key features of the EES were an annual coordination and monitoring of national employment policies based on Member States' commitments to establish a set of common objectives and targets. To its supporters, the EES ensured that the objective of a high level of employment saw the same political importance as the (other) macroeconomic objectives of growth and economic stability.

The EES added a dimension. The existing legislative pillar was about minimum standards, the EES was about policy directions, from stimulating employability to considering labour taxation. Employment coordination at the EU level meant – for the first time – a structured and systematic debate about labour markets at the EU level, and working towards and arriving at agreement about a common agenda and policy directions. This is a prerequisite if one is to set common goals such as the employment guidelines and arrive at a more integrated social-economic

agenda. It specifically saw three components: building of, and agreement on a common set of indicators to track progress/problems; understanding different systems, e.g., the role of social dialogue; and, understanding different policy outcomes and the possibilities/limits of the transfer of good practices.

As the EES was becoming more and more embedded in European and national employment policy-making, the Commission launched an ambitious, wide-ranging programme to make the EU "the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion." This programme was agreed at the Lisbon European Council in March 2000 giving rise to the Lisbon Strategy (Lisbon European Council 2000). With its commitment to "more and better jobs," European employment policies and in particular the EES were to be a key part. The employment guidelines, discussed with and adopted by Member States each year, contained detailed guidance as to how they would obtain more and better jobs. The employment guidelines for 2001 saw 18 separate detailed aspects of how to do this including some quantitative targets – key aspects of the Lisbon Strategy – such as an employment rate of 70% overall including at least 60% for women and 50% for older workers by 2010 (European Communities 2001).

The Kok Report "Jobs, Jobs, Jobs" Begats Flexicurity

Though subject to annual revisions, Member States displayed increasing dissatisfaction with the way the EES was being run. It was seen as too bureaucratic not sufficiently focused on the key challenges and not always coherent with the Broad Economic Policy Guidelines (BEPGs). In March 2003 the European Council invited the Commission to establish a European Employment Taskforce to be headed by former Dutch Prime Minister Wim Kok. It was charged with carrying out an independent in-depth examination of key employment-related challenges and to identify practical reform measures. It reported in November 2003 (Kok 2003) and its key findings were that increased efforts needed to be made to boost productivity and employment in Europe. Four priority actions were needed: increasing adaptability of workers and entrepreneurs; attracting more people to the labour market; investing more and more effectively in human capital; and, ensuring effective implementation of reforms through better governance. The first three actions stemmed from a hard-headed analysis of the economic workings of Europe's labour markets. It provided the analytical and political underpinnings of the major policy priority of the EES – enhancing flexicurity. This, with ups and downs linked with the economic and financial crisis, has remained a core element of European employment policies until today. Similarly, the employment guidelines became and have remained much more integrated with the BEPGs with finance ministers in recent years often the major champions of flexicurity.

Getting Member States to agree on the principles of flexicurity was not easy. The trade unions were suspicious and fearful believing most of the labour market reforms would increase flexibility with reduced employment protection legislation (EPL) and little in the way of increasing employment security. One reason for such doubts was that flexibilisation usually cost the public purse nothing while investing in ALMPs, more comprehensive training or better social protection was expensive at least in the short term for public budgets and finance ministries (that in any case often seemed to believe the less EPL the better).While the Commission was trying to convince Member States of the benefits of flexicurity, the European Trade Union Institute noted, in early 2007, "the permanent place flexicurity is acquiring in the Commission's employment policy. . .. There has been little attention paid to the broader EES which now seems to stand in the shadow of the more confused flexicurity approach" (Keune and Jepsen 2007).

Employment ministers agreed the principles of flexicurity in late 2007 (Council 2007). It did dominate the EES until the crisis, but the aspects of lifelong learning and skills saw a particular emphasis in 2008/9. This manifested itself as a determination to equip workers displaced by changing technology and globalisation with the light (new) skills for the new jobs the EU would need to create. Human capital investment had

long been seen as a vital component of more and better jobs. Workers were more productive but also more adaptable with a good skills basis. Getting new skills for new jobs was also something both employers and trade unions could agree on, one aspect of flexicurity that saw relatively non-conflictual social dialogue. The Commission's Communication highlighting the importance of and means to achieve new skills for new jobs (European Commission 2008a) was published in late 2008 as it was becoming clear that the EU was not seeing a normal economic downturn but a major recession, originating and largely imported from the USA but now global in nature.

The Crisis and Stressing Short Time Working

The financial and economic crisis saw output fall quite dramatically in most European countries as banking systems imploded and export markets collapsed. Unemployment, with the usual lag, started to increase after 6 or 7 years of sustained falls across the EU. Most policy-makers thought a large part of this decline was due to good employment policies being implemented, sometime with politically difficult struggles. Germany, with its hard-won Hartz reforms, was seen as the leading example of what had needed to be done. The key features included *inter alia* more active support including vocational training for the unemployed, "mini-jobs" with much lower social security contributions and less generous unemployment benefits. These increased the adaptability of labour markets by increasing flexibility of jobs but also security for employment As the crisis deepened, it was tackled by most Member States with conventional Keynesian increased fiscal spending under the Commission's European Economic Recovery Plan. It proposed a coordinated economic stimulus of 11/2% of EU GDP with budgetary spending supposed to be timely, targeted and temporary (European Commission 2008b). Nevertheless, EU unemployment rose steadily although less fast than in the USA. Nearly all Member States saw a rise but some much more than others and such rises were often uncorrelated to the amount that output had fallen. Germany saw a big fall in GDP as exports crashed but little rise in unemployment. Spain, with a similar GDP decline and caused first and foremost by its construction bubble bursting, saw a much steeper increase in joblessness. This divergent response of labour markets to the crisis continued into aftermath and suggested that although macroeconomic conditions were very determinant for the levels of unemployment, labour markets and employment policies that affected their working could also play a major role.

During the crisis the EES had focused on braking the rise in unemployment through the use of short time working schemes in which the state paid part of the income of workers put on shorter hours. Germany was a major user of such schemes and Belgium, France, the Netherlands and Italy also used them significantly. Not only did it ensure better incomes and thus more aggregate demand in the economy (and political support) compared with more workers simply being put on the dole, but keeping workers at work and providing training in some of the hours they were not working would reduce the risk of permanent loss of skills and hysteresis. The German "Kurzarbeit" scheme was seen as a useful complement to the Hartz reforms and fully in line with the flexicurity paradigm. (Short time working was a form of internal flexicurity.) The OECD and Commission worked together on identifying the best employment policies to confront what many were already calling the Great Recession (see European Commission 2010).

Prolonged Crisis, the Europe 2020 Strategy and a Focus on Youth and the Long-Term Unemployed

The crisis and a new Commission wanting to make its own mark rendered the Lisbon Strategy obsolete in its final months. It was superseded by a more focused programme, the Europe 2020 Strategy to achieve smart, sustainable and inclusive growth with European and national targets in five domains: research and development expenditure; environmental improvement; employment; educational attainment; and poverty reduction. The last three areas found themselves in the four employment guidelines the Council adopted in October 2010; they also contained the EU targets for 2020: an employment rate of 75% for men and

women aged 20–64; reducing early school leavers to 10% and raising the share of 30–34 years with tertiary or equivalent level education to 40%; and reducing by 20 million the number of people at risk of poverty. The annual cycle of reporting, guidelines and recommendations in the EES, already merged with that of the BEPGs in the Lisbon Strategy, was subsumed into the European Semester process that the Europe 2020 Strategy gave rise to. The Joint Employment Report became an annex of the Annual Growth Survey the Commission published each Autumn to launch it.

Unemployment continued rising and then plateaued in late 2010. The crisis appeared to be over as output recovered. Germany and others phased out their short time working arrangements. Interest rates were increased by the ECB in 2011 as unemployment was just beginning to decline. But financial markets did not believe all of Europe's economy was back as the road to prosperity. Greece and other EMU countries on the geographic periphery of Europe saw a growing reluctance to be lent money through government bonds. Output fell once again and the EU entered its double dip crisis. The renewed recession in 2011 threatened to undo the Europe 2020 Strategy almost before it had started. Employment rates fell rather than rose and the number of people at risk of poverty increased rather than decreased. As unemployment again began to rise in the EU, and especially in the EMU periphery, it was the labour market group that had suffered most which began to become the new policy priority – young people. Youth unemployment, usually defined as the unemployment rate of those aged 15–24, had been higher than adult unemployment for several decades. But as the crisis had first unfolded rates in the EU had risen from 15% in 2008 to 23% in 2010 and were now going higher particularly in Greece, Portugal, Spain, Ireland and Italy.

As it had done with flexicurity, the Commission used devices or instruments outside the employment guidelines (even if these came to reflect the main political thrust contained in them) to prioritise action for the EES. Policy guidance was issued in the form of packages. In April 2012 the Employment Package focused substantially on the demand side of job creation just as fiscal consolidation was having an opposite effect. The policy prescriptions included reducing taxes on labour and supply business start-ups. It also underlined the need for greater involvement of the social partners in setting EU priorities. In December 2012 the Commission sought to address the persistent, and in some Member States still rising, youth unemployment. It proposed a European Youth Guarantee, which was adopted as a Council Recommendation in April 2013, in which all young people under the age of 25 would receive a good quality offer of employment, continued education or an apprenticeship of traineeship within 4 months of them leaving formal education or becoming unemployed (European Commission 2012). And, exceptionally, extra funding was found to the tune of six billion Euros to be used via the European Social Fund to help the most affected Member States finance such measures.

Unemployment in the EU, both overall and for young people, actually peeked in 2014. The crisis appeared to be over but in September 2015 a third package was prepared by the Commission to deal with one potentially long lasting or even permanent consequence; the high numbers of unemployed people who remained unemployed more than 1 year – the long-term unemployed (European Commission 2015). It varied widely across the EU but was as high as 19.5% of the active population in Greece. Inspired by the Youth Guarantee, the key element was to ensure all those who had been unemployed for at least 18 months were registered with national employment services, received an individual in-depth assessment of their employability including skills needs and signed up into a job integration agreement. Member States agreed to the package in February 2016 even though there was no additional funding proposed to help them facilitate the re-entry of the long-term unemployment into employment. Perhaps the recovering European economies were supposed to do this without significant extra investment in ALMPs for the long-term unemployed. But those Member States with the highest long-term employment rates tended to be those under the most pressure to continue fiscal consolidation and thus had least resources to finance additional ALMPs themselves.

What Have European Employment Policies Achieved?

In early 2017 EU unemployment is just above 8% while the employment rate is around 70% for 20–64 year olds: the former is a little higher than the best figures just before the crisis while the latter is fractionally better than that in 2008. These averages hide large variations among the Member States with some still seeing very high levels of overall unemployment, youth unemployment and long-term unemployment. As the 2017 Joint Employment Report notes: "unemployment, youth unemployment and poverty levels remain far too high in many parts of Europe; labour market and social outcomes vary by gender, age and education; income inequality remains high in many EU countries with negative implications for economic output and inclusive and sustainable growth" (JER 2017).

In many ways the challenges confronting employment policy in Europe and the recommended measures to address them are not very different from those facing policy-makers when the EES was first launched 20 years ago. One clear difference is the greater emphasis put on poverty, social protection and inequality, issues largely absent for the first employment guidelines. These concerns have been an increasing part of the EES within the Europe 2020 strategy and its explicit poverty target of reducing those at risk of poverty by 20 million. They also reflect the legacy of the crisis which saw number of those at risk of poverty rise sharply rather than fall. Though new to the EES, these concerns had been evident in the social OMCs but these were far from fully integrated into the Europe 2020 strategy.

With the figures for employment and unemployment in the EU on average at best little better than 10 years before, and the challenges identified remaining quite similar, it seems difficult to say 20 years of the EES was a resounding success. Nevertheless, many would say it is unfair to measure its achievements looking at these data; they would say that it was the crisis and macro policy responses to it, especially the fiscal consolidation programmes, which overwhelmed the supply side changes employment policies could have brought about. Others would point to the apparent success stories of individual countries which weathered the crisis reasonably well, either by implementing EES-advocated reforms before it such as Germany or already had policies and institutions in place from the beginning along the lines of the flexicurity paradigm such as Sweden, Denmark, Austria and the Netherlands. There is some independent evidence that up until 2008, the EES led to convergence and improved employment outcomes over and above long-term or international trends (Van Rie 2012).

Poland, the Czech Republic and the Baltic States also have well-performing labour markets today at least when looked at with data for employment and unemployment rates. They are doing much better than they were in the late 1990s and it could be claimed that it was preparing for EU and thus EES membership, and as of 2004 being full members, that has played a significant role in their success stories. The critics of the EES would retort that it had little or nothing to do with this and all to do with their integration into German-led industrial supply chains and sound macroeconomic policies. The debates about the desired flexibility of labour markets continue to be similar to those when the EES was set up. It would be fair to say that the jury is still out on how effective the EES or employment policies in Europe have been. Similarly, there is no agreement on what difference the EES itself has made or whether the apparent successes in labour market performance in individual Member States because of policy reforms would have happened in any case through national pressures to raise economic performance. How much of the Hartz reforms was inspired by what the EES guidelines were advocating or did the EES take up Germany's reform model? Were the reforms in Spain and Italy in the wake of the crisis largely inspired by or even done through European guidance and examples from other countries or largely internally driven and designed? To end on a more positive note, there are very few people who say that employment policy coordination, including Council adopted Recommendations, at a European level is bad and that it should be left entirely to Member States. Even the UK which is

leaving the EU to "regain control" did not object most of the time to being in the EES, stressed the usefulness of learning from others and nearly always played a full part in developing it.

Workers' mobility was and is one of the four fundamental freedoms, together with goods, services and capital, underpinning the original common market. European employment policies have sought to promote it to enhance productivity and employment across the EU. During the crisis it was seen as a significant means to reduce divergence; the unemployed from countries with high unemployment were actively encouraged to seek work in countries with much lower rates. But recent times have seen growing concern that this puts undue pressures on host countries. Combined with a surge of refugees, both political and economic, concentrated in just a few Member States, and growing support for nationalist political parties, enhanced labour mobility is little seen and less talked about as a priority for the EES. The same Joint Employment Report for 2017 identifies four priorities for the EES to promote the creation of quality jobs: removing barriers to labour market participation, tackling labour market segmentation and undeclared work, ensuring that social protection systems provide adequate income support and ensuring all people had access to enabling services while transactions into employment and making work pay are encouraged. These can be considered its work programme for the next few years.

Acknowledgments The author would like to thank Carola Bouton, Lieselotte Fürst, Cristina Martinez Fernandez and Tim Van Rie for valuable contributions to this entry.

This is written in a personal capacity and does not necessarily reflect the views of the European Commission.

Bibliography

Council of the European Union. 2007. Towards common principles of flexicurity – Council conclusions, 6 Dec 2007.

Delors, J. 1993. *Growth, competitiveness and employment, white paper*. European Commission, Brussels.

EEC. 1957. Treaty establishing the European economic community. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3Axy0023.

European Commission. 2008a. New skills for new jobs. COM (2008) 868.

European Commission. 2008b. A European economic recovery plan. COM (2008) 800.

European Commission. 2010. Short time working arrangements as response to cyclical fluctuations. http://ec.europa.eu/social/main.jsp?langId=en&catId=89&newsId=843&furtherNews=yes.

European Commission. 2012. Moving youth into employment. COM (2012) 727.

European Commission. 2015. The integration of the long-term unemployed into the labour market. COM (2015) 462.

European Communities. 1972. Bulletin of the European Communities. October No 10. Statement from the Paris Summit, 14–26. http://www.cvce.eu/content/publication/1999/1/1/b1dd3d57-5f31-4796-85c3-cfd2210d6901/publishable_en.pdf.

European Communities. 2001. Official Journal 24/01/01, 18–26.

Goetschy, J. 1999. The European employment strategy: Genesis and development. *European Journal of Industrial Relations* 5 (2): 117–138.

JER. 2017. Joint employment report from the commission and the council, council of the European Union, 3 March.

Keune, M., and M. Jepsen. 2007. ETUI. WP 2007.01.

Kok, W. 2003. Jobs, jobs, jobs: creating more employment in Europe. Report by the Employment Taskforce.

Layard, R., R. Jackman, and S. Nickell. 1991. *Unemployment*. Oxford: Oxford University Press.

Lisbon European Council. 2000. Presidency conclusions. 23–24 Mar 2000.

OECD. 1997. Press communiqué at end of 1997 Ministerial meeting.

Rhodes, M. 2005. Employment Policy: Between efficacy and experimentation. In Wallace, H., Wallace, W. and Pollack M. (eds) Policy Making in the European Union. Oxford: Oxford University Press.

TEU. 1997. Treaty of Amsterdam amending the treaty of the European Union. Official Journal of the European Communities C340, 10 November.

Van Rie, T., and I. Marx. 2012. The European Union at work? The European employment strategy from crisis to crisis. *Journal of Common Market Studies* 50 (2): 335–356.

European Labour Markets

Giuseppe Bertola

Abstract

Labour taxes and subsidies, collective wage bargaining, and employment protection

legislation affect labour market outcomes in European countries more strongly than in other advanced countries. This article outlines theoretical approaches to their motivation and consequences and reviews empirical insights from comparative crosscountry studies of how employment, unemployment, and wage dynamics are shaped by the interaction between institutions, macroeconomic developments, and structural features.

Keywords

Active labour market policies; Centralized wage determination; Efficiency wages; Employment protection legislation; European labour markets; Health insurance; Inflation; Job creation; Labour market search; Labour supply; Labour taxes; Minimum wages; Pensions; Structural change; Tax wedges; Training; Unemployment insurance; Unemployment subsidies; Wage floors; Wage rigidities; Women's work and wages

JEL Classifications

J0

European labour markets, especially those of Continental countries, are characterized by more unionized wage setting and more stringent regulation of employment relationships than those of other OECD countries. Within that group of advanced countries, their unemployment rates used to be relatively low, and became very high. Around 1970, the unemployment rate was approximately 3.1 per cent in the OECD aggregate and five per cent in the United States, but the unemployment rate hardly exceeded four per cent in any European country. In the aggregate of 11 core European Union countries that later adopted the euro at its inception, unemployment stood at only 2.2 per cent in 1970. It then rose rapidly, exceeding ten per cent in 1984 and hovering around 12 per cent in the second half of the 1990s, while both the United States and the OECD aggregate unemployment rates fluctuated between four per cent and nine per cent.

The wide variety of labour market developments over the last quarter of the 20th century has motivated extensive modelling efforts and comparative empirical studies of institutional features' motivation and effects. This article reviews the roles of institutions, shocks, and structural change in shaping aggregate and disaggregate labour market outcomes.

Labour Market Policies

To illustrate the spirit of more general approaches to the relevant issues, it is useful to focus initially on the simplest models and the best understood labour market institutions (Prescott 2004). Consider inverse demand and supply functions

$$w^d = a(l), \quad w^s = s(l), \tag{1}$$

where l denotes log employment and w^s and w^d denote log wage rates. The wage w^* and employment l^* that equate supply and demand satisfy the condition

$$s(l^*) = a(l^*) = w^* \tag{2}$$

in static competitive equilibrium. As the simplest example of how institutions can change this outcome, consider a labour income tax that, inserting a wedge τ between employers' labour costs and workers' take home pay, changes the equilibrium condition to

$$s(\underline{l}) = a(\underline{l}) - \tau \tag{3}$$

and lowers employment by about

$$\underline{l} - l^* \approx -\tau/(\eta + \varepsilon) \tag{4}$$

where $s'(l) = \varepsilon \geq 0$ and $a'(l) = -\eta$, $0 < \eta < 1$. It is also simple to characterize formally the effects of binding legal or contractual minimum wage levels. If the wage is $\underline{w} > w^*$, the employment levels corresponding to $\underline{w}$ on the supply and demand curves are defined by $s(L) = \underline{w}$ and $a(\underline{l}) = \underline{w}$, and differ by the number

$$L - \underline{l} \approx (\underline{w} - w^*)(\varepsilon + \eta)/\varepsilon\eta \tag{5}$$

of unemployed workers, who would be willing to work at the going wage but cannot obtain employment.

From this simple perspective it is obvious that differences in taxation and wage floors may explain cross-country differences in employment and unemployment. Qualitatively similar insights can be derived in the context of more complex and realistic models of unemployment, and can be applied to other institutions. When unemployment is due to matching frictions, efficiency wages and other imperfect allocation mechanisms, taxes and wage rigidities can affect search efforts and equilibrium employment and unemployment, which are affected in turn by the market's structure (such as the extent of mismatch between workers' qualifications and vacancies) and by other institutional features (such as the scope and efficiency of employment agencies). In both competitive and frictional models of the labour market, benefits paid to out-of-work individuals can affect labour supply and search effort, and there can be similar effects from less visible policy aspects, such as the availability of public-sector employment opportunities at favourable wage–effort ratios (Algan et al. 2002).

At the same time as it offers obvious explanations for labour market outcomes, institutional variation raises the less obvious issues of why institutions should be as different across countries as they are observed to be, and of how their configuration and impact may depend on structural labour market features.

The relevance of distributional issues and of market imperfections can explain some of the labour market institutions' heterogeneity. The equilibrium condition (1) efficiently equates employed labour's marginal productivity with its non-employment opportunity cost, and distorting this outcome reduces the welfare of a perfectly competitive economy's representative individual. If workers disregard non-labour income, however, their total surplus can be increased by trading lower employment against higher pay along downward-sloping labour demand curves such as (2). It is maximized when the wage exceeds the marginal opportunity cost of employment by a monopolistic markup factor, and employment is set at a level $\underline{l}$ such that

$$a(\underline{l}) - s(\underline{l}) = \log[1/(1-\eta)] \approx \eta. \qquad (6)$$

All workers' welfare can be increased if the higher wages earned by those who are employed more than compensate for the labour income lost by those who would be employed at the competitive wage. Such compensation may take place within families, or over individual lifetimes, and can also be explicit if the revenue raised by employment taxes is spent subsidizing non-employed individuals.

Institutions that decrease employment and increase labour costs can be rationalized recognizing that they affect not only the amount of production but also its distribution across heterogeneous individuals, and that markets (especially financial markets) are not perfect in real-life economies. Higher wages and lower employment can benefit workers who have negligible non-labour income, and households' limited access to formal financial markets can rationalize collectively administered risk-sharing schemes (Agell 2002). In European countries, legislation meant to endow workers with some bargaining power and to insure them against health, unemployment and old-age hazards was introduced at times of actual or feared social unrest, in Bismarck's industrializing Germany or in Lord Beveridge's post-war United Kingdom. In principle, it can be efficient to try to provide insurance through mandatory government schemes when information and legal enforcement problems make it difficult for private markets to do so. But public schemes are not immune from such problems, and tend to reduce employment as, for example, recipients of unemployment subsidies reduce work effort. Such efficiency losses are more easily affordable by richer societies, and Europe's fast and stable post-war growth was unsurprisingly accompanied by development of increasingly extensive legislation and co-decision powers by unions. By the early 1970s, the institutional structure of labour markets was distinctively different not only across the United States and Europe as a whole, but also across countries

within Europe, where labour market policies play different roles in different welfare state models (Bertola et al. 2001). In Nordic countries, a tradition of full employment and universal welfare is based on generous unemployment benefits and a very important role for active labour market policies (including job creation in the public sector). The Bismarckian model of Continental countries such as France and Germany features centralized wage determination and stringent employment protection legislation, and contributory pension, health, and unemployment insurance programmes. The Beveridgian model of the United Kingdom and other Anglo-Saxon countries features social assistance safety financed by general taxation and comparatively light regulation of wage determination and employment relationships.

The Dynamics of European Labour Market Outcomes

Even though relief from the need to work should in general reduce employment, until the 1970s, and even in the aftermath of the late 1960s period of worker unrest, increasingly generous pro-worker institutions coexisted in Europe with low unemployment rates; much lower, in fact, than in the comparatively unregulated United States. The first oil shock and the following decades of slower growth saw the inception and persistence of high unemployment in most European countries, and increasing attention to the effect of institutions on labour market performance. If wages are preset, shocks can cause employment and unemployment fluctuations, the size and persistence of which depends on the extent of *ex post* wage flexibility and on the character of wage bargaining. Nominal shocks are a more relevant source of real wage misalignments and unemployment in labour markets with more pervasive and longer-term collective wage contracts. Conversely, real wages react more promptly to productivity shocks or growth slowdowns if bargaining parties are in a better position to take into account their employment implications. Reactions to country-wide shocks are quicker, and the unemployment consequences of such shocks less severe, when wage bargaining is more centralized and better coordinated across industries (Calmfors and Driffill 1988).

This can explain why unemployment began to increase, more or less sharply, when in the 1970s European countries were hit by oil shocks and other macroeconomic developments that reduced the amount of labour demanded at any given wage. Inflation and output dynamics subsequently appear to drive European unemployment fluctuations around a natural level that, after having raised sharply until the early 1980s, has remained essentially flat since the mid-1980s (Blanchard 2006). The prolonged upward trend and the resilience of high unemployment levels naturally draw attention to non-cyclical, structural aspects of labour market dynamics. Wage floors can prevent underbidding by the unemployed of Eq. (5), but it is difficult for that static relationship to explain why, in the absence of institutional changes that would further increase unions' wage-setting power, unemployment remained high in the aftermath of the 1970s crises.

A more suitable dynamic perspective is offered by models where labour demand shocks can permanently affect the link between wages and outside options, for example because job losers no longer have a say in wage determination, or because replacement of employed workers would entail large turnover costs (Lindbeck and Snower 1988). The persistence of employment and unemployment dynamics, however, is in fact influenced not only by limited wage-setting flexibility but also by regulatory constraints on hiring and firing. In European countries, employment protection legislation (EPL) typically requires that the reasons for individual dismissals be stated by employers and subject to court appeal, and that collective dismissals be conditional on administrative procedures involving formal negotiations with workers' organizations and with local or national authorities.

Such provisions do have the intended effect of 'protecting' jobs at times of declining labour demand, when firing costs smooth out job losses and reduce downward wage pressure. Just because such a situation is costly for employers,

however, it is optimal for them to refrain from hiring in upturns, so as to reduce the desirability of labour shedding in downturns. In terms of simple demand-and-supply relationships such as those introduced above, the marginal productivity of labour should be lower than the wage when employment is declining and firing a marginal worker entails firing costs as well as wage-cost savings, but it should symmetrically be higher than the wage when employment is increasing, and the marginal worker's costs include expected future firing costs as well as the current wage. Thus, the implications of EPL are similar to those of labour taxes for expanding firms, and to those of employment subsidies for downsizing firms. If employment fluctuations are efficient in laissez-faire, EPL obviously reduces production and profits. Unlike labour taxes, however, it does do not do so by reducing employment on average (Bentolila and Bertola 1990), because its contrasting effects on employers' propensity to hire and fire reduce employment volatility but affect its average level ambiguously. Empirically, in fact, there is no convincing evidence of any relationship between EPL and the employment or unemployment level. As discussed in some detail below, correlations have to be treated with caution in this context, but more stringent EPL is associated with more stable aggregate employment paths and with longer unemployment durations within the pool of unemployed workers (Bertola 1999). There is also some evidence that EPL affects the demographic composition of employment and unemployment – as it should in theory, since it reduces job finding rates for young job market entrants and female workers with intermittent labour force participation at the same time as it reduces job-loss rates for mature workers.

Another important related difference across labour markets pertains to the extent and character of wage inequality. Earnings are typically less dispersed in Europe than in other advanced countries. The extent of underlying heterogeneity in workers' characteristics is an important determinant of earnings dispersion, but institutional wage-setting constraints also appear very relevant, both theoretically and empirically. While centralized bargaining may be better able to coordinate reactions to aggregate shocks, it tends to result in less detailed, more homogenous wage structures across firms, sectors, regions and individuals. Similar wages for heterogeneous workers imply divergence of employment outcomes, for example across demographic groups (Kahn 2000) and across regions in Italy, Germany and Spain, where the uniformity of centrally bargained wages (and of other national institutions) tends to lower employment where labour is less productive. Empirically, relative wage variation appears to be heavily constrained in the same countries where EPL is most stringent (Bertola and Rogerson 1997). This is unsurprising, because quantitative firing restrictions could hardly be binding if, in the face of negative labour demand shocks, wages could fall so as to make stable employment profitable, or to induce voluntary quits. Across countries, the combination of wage and quantity rigidities indeed appears to protect employed workers from labour income volatility, as individuals enjoy more stable wages and longer tenures.

At the aggregate level, the role of institutions in shaping heterogeneous dynamics across labour markets is not as immediately apparent. Institutions vary widely across countries but, within each country, they are much more stable than unemployment, wage inequality and other labour market outcome variables. As discussed above, however, wage-setting institutions can shape an economy's reaction to aggregate shocks. More generally, the same dynamic developments can produce very different employment and wage outcomes in countries with different (albeit stable) institutions. This can explain why, in the 1970s and 1980s, countries with more extensively regulated labour markets experienced more pronounced unemployment increases in the aftermath of similar productivity, inflation and wage shocks (Blanchard and Wolfers 2000). Empirically, in fact, the forces that interact with labour market institutions in driving dynamic trajectories can be almost equally well represented by period-specific dummy variables as by observable macroeconomic variables, which tend to behave rather similarly over time across industrialized countries. Thus, the evidence can be consistent with a role for common structural trends rather than for country-specific shocks.

For example, the relationship between country-specific labour market institutions and unemployment and wage dispersion dynamics can be interpreted in the light of skill-biased technological progress trends, or of increasing opportunities for advanced countries to import unskilled labour-intensive goods and export skillintensive ones. Over the last three decades of the 20th century unemployment displayed a trend increase in Continental European countries but remained trendless in the United States and other Anglo-Saxon countries, while earnings inequality remained stable (or even declined) in the former group of countries but trended upward in the latter. If technological progress or international trade increase laissez-faire wage inequality, they also increase the relevance of wage floors: if in European countries low wages cannot decline, employment of unskilled workers must decline (Krugman 1994). Similar insights into the changing implications of unchanging institutions can be gained by considering other structural aspects. More intense product market competition, as implied by Europe's economic integration process and by more general globalization trends, increases the elasticity of labour demand.

In the context of the simple example above, a smaller η implies larger employment losses from any given tax wedge in Eq. (4), and higher unemployment from any given wage floor in Eq. (5). In more complex dynamic models, if reallocation towards higher-paying jobs is costly, then institutions that tend to prevent wage inequality and restrict mobility have sharper implications for employment and unemployment when more volatile shocks affect labour demand (Ljungqvist and Sargent 1998).

Structural change can magnify the unemployment and employment effects of institutions meant to redistribute income and remedy financial market imperfection, or it can make them redundant (for example, because financial market development makes labour income fluctuations less problematic). Then, institutions should be reformed. In the simple formal framework above, the same smaller η that amplifies the negative employment implications of given institutions also calls for a smaller markup in Eq. (6). And, in reality, policy frameworks introduced in the 1990s, such as those recommended by the OECD Jobs Study (OECD 1994) and by the European Union's Lisbon Strategy (Council of the European Union 2000), de-emphasize income support for job seekers and job losers in favour of job creation spurred by wage and employment flexibility, and the role of training and other active labour market policies aimed at bringing workers' productivity in line with wage aspirations.

Reforms are at least partly motivated by better theoretical and empirical understanding of the effects of labour market institutions. But while it is in principle obvious that institutional interference can be responsible for high unemployment and low employment, just because such effects depend on potentially heterogeneous structural parameters, that it is hard to assess their impact in data where many relevant confounding factors cannot be controlled. Simple correlation can be very misleading. For example, a negative cross-country correlation between EPL and employment rates is fully accounted for by low female employment–population ratios in southern Europe (Nickell 1997), while effects on prime-age male employment rates tend to be positive. Both policies and outcomes can jointly respond to underlying cultural differences in this and other cases, and it is difficult to obtain reliable estimates from cross-sectional relationships between institutions and outcomes (Baker et al. 2005). More articulate and robust insights may be obtained from specifications where time-series variation and interactions play important roles (Bassanini and Duval 2006). As the time dimension of available data increases, however, it will be increasingly important when interpreting time-series evidence to focus on the economics and politics of reform processes rather than on institutions at each point in time (Saint-Paul 2000), and to be aware of plausible channels of institutional endogeneity. If shocks or structural changes make job loss more or less likely or trigger painful changes in the generosity of unemployment insurance or in the stringency of employment protection legislation, for example, the correlation between such institutions and employment performances may be largely spurious. The wide and changing

variety of labour market policies across countries offers opportunities to try to disentangle their effects in increasingly available disaggregated data, at the same time as it makes it necessary to take into account the many important and related respects, besides labour market structure, in which countries differ.

See Also

- ▶ European Monetary Union
- ▶ Globalization and the Welfare State
- ▶ Phillips Curve
- ▶ Phillips Curve (New Views)
- ▶ Skill-Biased Technical Change
- ▶ Unemployment

Bibliography

Agell, J. 2002. On the determinants of labour market institutions: Rent seeking vs. social insurance. *German Economic Review* 3: 107–135.

Algan, Y., P. Cahuc, and A. Zylberberg. 2002. Public employment and labour market performance. *Economic Policy* 34: 9–65.

Baker, D., A. Glyn, D. Howell, and J. Schmitt. 2005. Labour market institutions and unemployment: A critical assessment of the cross-country evidence. In *Fighting unemployment: The limits of free market orthodoxy*, ed. D. Howell. Oxford: Oxford University Press.

Bassanini, A., and R. Duval. 2006. Employment patterns in OECD countries: Reassessing the role of policies and institutions. Economics Department Working Paper No. 486; Social, Employment and Migration Working Paper No. 35. Paris: OECD.

Bentolila, S., and G. Bertola. 1990. Firing costs and labour demand: How bad is Eurosclerosis? *Review of Economic Studies* 57: 381–402.

Bertola, G. 1999. Microeconomic perspectives on aggregate labour markets. In *Handbook of labour economics*, ed. O. Ashenfelter and D. Card, Vol. 3C. Amsterdam: North-Holland.

Bertola, G., J.F. Jimeno, R. Marimon, and C. Pissarides. 2001. Welfare systems and labour markets in Europe: What convergence before and after EMU? In *Welfare and employment in a United Europe*, ed. G. Bertola, T. Boeri, and G. Nicoletti. Cambridge: MIT Press.

Bertola, G., and R. Rogerson. 1997. Institutions and labour reallocation. *European Economic Review* 41: 1147–1171.

Blanchard, O.J. 2006. European unemployment: The evolution of facts and ideas. *Economic Policy* 45: 7–59.

Blanchard, O.J., and J. Wolfers. 2000. The role of shocks and institutions in the rise of European unemployment: the aggregate evidence. *Economic Journal* 110: C1–C33.

Calmfors, L., and J. Driffill. 1988. Centralization of wage bargaining. *Economic Policy* 3: 14–61.

Council of the European Union. 2000. *Conclusions of the Lisbon European Council*, Council of the European Union SN 100/00, 23–24 March.

Kahn, L.M. 2000. Wage inequality, collective bargaining and relative employment 1985–94: Evidence from 15 OECD countries. *The Review of Economics and Statistics* 82: 564–579.

Krugman, P. 1994. Past and prospective causes of high unemployment. *Federal Reserve Bank of Kansas City Economic Review* 1994(4): 23–43.

Lindbeck, A., and D.J. Snower. 1988. *The insider–outsider theory of employment and unemployment*. Cambridge: MIT Press.

Ljungqvist, L., and T.J. Sargent. 1998. The European unemployment dilemma. *Journal of Political Economy* 106: 514–550.

Nickell, S. 1997. Unemployment and labour market rigidities: Europe versus North America. *Journal of Economic Perspectives* 11(3): 55–74.

OECD (Organisation for Economic Co-operation and Development). 1994. *The OECD jobs study: Evidence and explanations*. Paris: OECD.

Prescott, E.C. 2004. Why do Americans work so much more than Europeans? *Federal Reserve Bank of Minneapolis Quarterly Review* 28(1): 2–13.

Saint-Paul, G. 2000. *The political economy of labour market institutions*. Oxford/New York: Oxford University Press.

European Monetary Integration

David G. Mayes

Abstract

This article explores the development of Economic and Monetary Union in Europe from the Second World War through to the end of 2010. It concentrates primarily on the earlier part of the process and contrasts what has been implemented since 1999 with its antecedents and with the prescriptions suggested by economic theory. It covers the work of the Werner Committee, the European Monetary System (including its Exchange Rate Mechanism), the

provisions of the Maastricht Treaty, the Stability and Growth Pact and their implementation.

Keywords
Delors Committee; Economic and Monetary Union; Europe; European Monetary System; Exchange Rate Mechanism; Integration; Maastricht Treaty; Monetary union; Stability and Growth Pact; Werner Report

JEL Classifications
E42; F15; N14

It is easy with the benefit of hindsight to treat the path to the present euro area as if it was an inevitable and carefully planned process. In practice it has been a series of decisions taken in the light of the broad goal of closer integration on the one hand and the more immediate needs and concerns on the other. If this had not been the case then it is unlikely that the EU would have reached the difficulties at the time of writing (2011), where Greece, Ireland and Portugal have been forced into an emergency support programme from the EU and the IMF because markets regard their debt financing programmes as unsustainable within the euro area. Monetary union is not a topic which can somehow be treated separately from the rest of economic integration. Indeed, it is important to recall in the EU context that EMU stands for Economic and Monetary Union and not European Monetary Union. All the work on optimum currency areas makes it very clear that integration in product and labour markets as well as the existence of complementary structures and policy frameworks are essential if a monetary union is to work well. The process is clearly incomplete in the European environment. In many respects the process of closer integration in Europe has been opportunistic in the sense that integration has progressed in those dimensions that appeared tractable at the time.

In this article, therefore, I map the progress towards monetary union in Europe since the Second World War against the criteria which are required for a successful monetary union, so that the balance between the political decision-making and the economic requirements can be clear. The article is in four parts: monetary integration in the immediate post war period; the Werner Plan and the creation of the EMS (European Monetary System); the road to the euro area; and the development of monetary union. In that final section I relate the structure of the existing system to the problems encountered in 2010 and the real threats to the continuation of European monetary union in the form currently envisaged. These four phases cover the periods: 1944–1970; 1971–1992; 1992–2001; and 2002 onwards.

Monetary Integration in the Immediate Post-war Period

The European input to the design of the post-war international monetary system was largely undertaken by the UK, but the final form of the system agreed at Bretton Woods in 1944 was mainly US inspired and was centred on the USA. The monetary system in Eastern Europe was imposed by the Soviet Union. The Bretton Woods system itself was a reaction to the interwar experience and the problems of adjustment after the First World War. The immediate post-war experience had not been a happy one. Germany had dissolved into hyperinflation and for most other countries the intention had been to return to the stability of the gold standard. But the attempts to return at parities that were too high had imposed strains and a decade later the system in the largest economy, the USA, itself collapsed, resulting in the Great Depression. Countries looked to protect themselves on the one hand and tried to obtain competitive exchange rates on the other.

The Bretton Woods system offered stability by comparison. Exchange rates were to be fixed but adjustable should parities become unsustainable. Fixity was not total, but a fluctuation of $\pm 1\%$ was permitted. While fluctuations of that size offer arbitrage opportunities, they are small enough to be neglected by commercial businesses in setting prices. Although there were some initial realignments, by and large the system worked rather well for 20 years and was the basis of the post-war recovery. The Bretton Woods agreement also led

to the setting up of the International Monetary Fund (IMF) which could use deposits from all of the member countries to provide temporary assistance to a country that was having balance of payments difficulties. This reflected the inherent asymmetry of the system. A country that is running surpluses has little problem maintaining its exchange rate at fixed parity. One which is running a deficit, on the other hand, will eventually run out of reserves. While other routes to restoring surpluses may be possible, the obvious route is to devalue. However, devaluations do not produce immediate surpluses; in fact they do the opposite in the short run as existing contracts are honoured before expenditure is switched towards domestic goods and exports in response to the change in relative prices – hence the need for temporary assistance by the IMF conditional on moving towards a new and sustainable long-term position at the new parity.

Thus by the mid-1960s, when strains began to emerge, the system of fixed but adjustable exchange rates was the peacetime system that European countries were accustomed to. The strains in the main came not from European countries but from the USA itself, which was the anchor of the system. All parities were expressed with respect to the US dollar, which was itself convertible into gold at $35/ounce giving a basis reminiscent of the gold standard. The US problems, to quite some extent, stemmed from the costs of financing the Vietnam war, and, once inflation began to take hold, a different anchor was needed.

No other country was in a position to take on the anchor role and hence the result was a drift into floating exchange rates, something only the Canadians had had any substantial experience of in the post-war period. What the countries of the European Community sought (as this was the Community of the Six up until 1971) was a route back to stability at least with respect to each other. The route chosen by the time of the Hague summit in December 1969 was to try to create a form of economic and monetary union in which exchange rates among the members would be fixed. The process was to be progressive, and a committee was set up under Pierre Werner, then Prime Minister of Luxembourg, to map out a plausible way forward. The Commission document on which the Hague resolution was largely based already incorporated the main ideas of the Werner Committee's ultimate proposals for an economic and monetary union by stages. It also drew on the Barre Report, issued in February 1969, which called for coordination to achieve medium term economic objectives.

The Werner Plan and the Creation of the EMS

The Werner Committee worked quickly, producing an interim report in May 1970 and a final report in October the same year. The proposals were adopted by the European Council in February 1971. The plan envisaged economic and monetary union being achieved in three phases over the course of the ensuing decade, so it would be completed in 1980. Such an economic and monetary union would include the four freedoms of movement of goods, services, labour and capital laid down in the Treaty of Rome as well as having 'a single monetary entity… characterized by the total and irreversible convertibility of currencies; the elimination of fluctuation margins of exchange rates… [and] the irrevocable fixing of parities' (Commission of the European Communities 1970, p. 10). While a single currency was not viewed as essential to this scheme, it was thought the best option. While it recognized the need for a single monetary policy at the Community level and agreement on medium-term economic objectives and coordination of shorter run economic policy, it expected the union to be relatively decentralized and for there to be no large Community budget or fiscal resources.

It was quite explicit about the form of the institutions for monetary policy 'The constitution of the Community system for the central banks could be based on organisms of the type of the Federal Reserve System' in the USA (Commission of the European Communities 1970, p. 13). The Community institution would take the interest rate and other monetary decisions and manage the Community's foreign exchange reserves. It is also clear that an organ of 'economic

government' was envisaged: 'While safeguarding the responsibilities proper to each it will be necessary to guarantee that the Community organ competent for economic policy and that dealing with monetary problems are aiming at the same objectives'. This approach to increasing coordination of economic and monetary policies had been signalled by the Commission as early as 1962 and the Committee of central bank governors had been set up in 1964 to assist the process.

The bold characteristics of the plan are summarized as (Commission of the European Communities 1970, p. 26):

> 'Economic and monetary union is an objective realizable in the course of the present decade provided only that the political will of the Member States to realize this objective. . . is present'
>
> 'Economic and monetary union means that the principal decisions of economic policy will be taken at Community level and therefore that the necessary powers will be transferred from the national plane to the Community plane. . .. The economic and monetary union thus appears as a leaven for the development of political union which in the long run it will be unable to do without.' (p. 26).

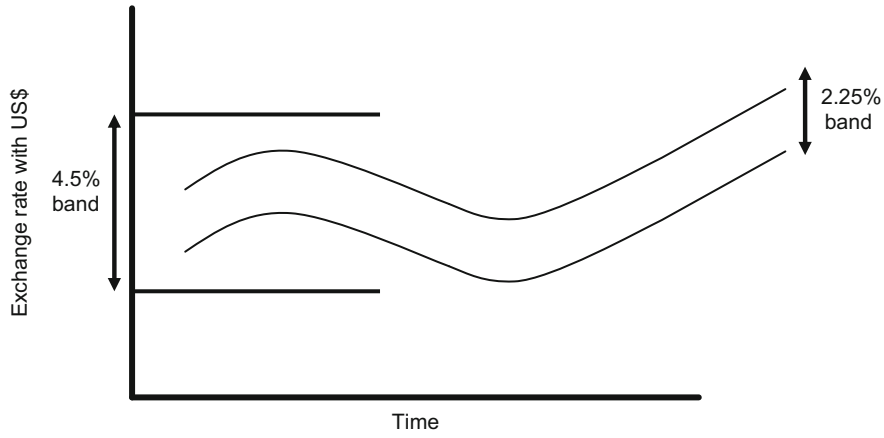
The requirements of the first stage were sweeping, going beyond what has been achieved in economic policy cooperation in the ensuing forty years, with a three-stage coordination of fiscal policy each year and a harmonization of tax instruments, organized in the first instance through the Council of economic and finance ministers. For monetary policy, 'From the start of the first stage, by way of experiment, the central banks acting in concert will limit de facto the fluctuations in the rates of exchange between their currencies to narrower margins than those resulting from the application of the margins in force for the dollar at the time of the adoption of the system. This objective will be achieved by concerted action in relation to the dollar'. The Committee of central bank governors would be required to make twice yearly reports on progress in developing the joint tools.

The progression would involve complete liberalization of capital movements and integration of financial markets. The final stage would however require a treaty revision in order to create the new institutions to replace coordinated national 'instruments' with 'community instruments' (Commission of the European Communities 1970, p. 24). These new institutions would include a European Fund for Monetary Cooperation under the control of the Governors of the central banks (Commission of the European Communities 1970, p. 25). Only the first stage had a timetable, three years.

The proposals did not really get off the ground as the dollar was effectively floated during 1971 and the best the countries could attempt was a joint float against the US dollar with limited fluctuation of each EC currency with respect to the others, as agreed by the central bank governors in April 1972. This mechanism was known as 'the snake' as illustrated in Fig. 1, where the shape of the joint movement over time might look somewhat like a snake. Under the snake the member states were required to keep their currencies within 2.25% of each other. In August 1971 the USA ended convertibility with gold and in December the Group of 10 largest countries made the Smithsonian Agreement, whereby the other countries appreciated with respect to the dollar and tried to keep their exchange rates within 62% of the new parity.

This proved difficult for many countries and the arrangements were abandoned in March 1973, when the system moved to floating exchange rates. By that time also the Community had been enlarged by the addition of Denmark, Ireland and the UK, the latter two having already started floating against the US dollar. This early period is known as the 'snake in the tunnel' as the Smithsonian Agreement permitted fluctuation of $\pm 4.5\%$, the edges of that band constituting the 'tunnel'. France, Ireland, Italy and the UK found remaining in the snake difficult and exited quite early with the impact of the first oil crisis. The system continued with Belgium, Luxembourg, Denmark, West Germany and the Netherlands until 1979 (the first two countries already having a common currency with parities fixed in 1944).

By 1978 the pressure to achieve more comprehensive exchange rate stability in the Community had risen to the point that the French President Valéry Giscard d'Estaing and the German Chancellor Helmut Schmidt proposed the creation of a



European Monetary Integration, Fig. 1 The snake

European Monetary System (EMS), following the suggestions of the then President of the European Commission, Roy Jenkins, the year before. The principle was agreed at the Bremen Council in the spring and the details agreed at the Brussels European Council in December. The arrangement came into force on 13 March 1979.

The key difference from the snake was that instead of being a dollar-based system it reflected purely intra-Community exchange rate fluctuations. At the heart of the system was the European Currency Unit (ECU) which was a weighted sum of the nine component currencies. (This followed on directly from European Unit of Account (EUA), a similarly weighted synthetic currency unit, which had been used internally in the Community in budgetary calculations.) The same $\pm 2.25\%$ bands were maintained, although for Italy a band of $\pm 6\%$ was instituted initially and the UK decided not to join in the exchange rate mechanism (ERM) part of the system. This last broke the link between the Irish punt and sterling. A European Monetary Cooperation Fund was also set up to ease ECU payments, with each country contributing 20% of its gold and foreign exchange reserves. Country weights in the ECU were recomputed at five year intervals, and Greece and then Portugal and Spain were included in the weights at the first opportunity after they joined the Community.

In the early years, fairly frequent alterations in parities were required, the first as early as September 1979 and then again in October 1981 and April 1982. By the time of the main problems with the EMS in 1992 there had been 12 realignments and none in the previous six years. The idea was that any country having difficulty maintaining its parity within the bands should start to intervene to keep within the limits when it had diverged by 75% of the permitted range. When it reached the edge of the range, intervention was supposed to be symmetric. However, in practice the burden of adjustment was placed largely on depreciating currency countries. Increasingly the system resembled a Deutschmark area, with West Germany setting interest rates for domestic purposes and the other countries having to follow suit in order to maintain their parities. As a result the EMS arrangements were amended by the Basle–Nyborg agreement of September 1987, which sought to encourage coordinated foreign exchange intervention and interest rate changes when a country approached the permitted limits of its parity. One important feature of the ERM was that changes in parities were intended to be decided by common consent and not by the country in difficulty unilaterally deciding where it would try to repeg its exchange rate. If that had been permitted there would have been a danger of competitive exchange rate changes and the system would not have been characterized by increasing stability.

Possibly the most significant feature of the EMS period was the development of the 'private' ECU, the issuing of ECU bonds and related

instruments. Large corporates and indeed governments found it cheaper and more efficient to issue ECU-denominated bonds rather than to issue bonds in a number of the component countries. It was not necessary to be a member of the system to do so. One of the larger issuers for example was the UK government, even though it was not participating in the ERM. Similarly the international financial institutions also had ECU offerings. It was thus clear that there was a demand for such a currency. Hence, while the system was designed in immediate terms to try to achieve greater exchange rate stability among the members, it acted as an important step towards a European financial market.

The apparent stability of the ERM in the period from 1986 to 1992 in fact covered up some of the underlying tensions in the system. Exchange rate stability is only possible if the countries involved are subject to common shocks, to which they react in a reasonably similar manner. By the end of the 1980s this was no longer the case, particularly with the collapse of the former Soviet Union and the unification of East and West Germany. Much as the collapse of the Bretton Woods system was led by problems in the base country, the USA, so the collapse of the ERM was to quite a large extent driven by problems that were specific to Germany. Unification resulted in a major fiscal challenge, which resulted in the need to raise interest rates. In countries that did not share this need, the appropriate strategy would have been to devalue, but with the ERM commitment this was not welcome. Strains therefore built up and the system began to fall apart as markets speculated against each of the most exposed countries in turn, thereby forcing the devaluations.

It is ironic that, as discussed in the next section, the EU was simultaneously trying to put in place the next steps towards monetary union, with the negotiation and ratification of the Maastricht Treaty. The rejection of the treaty in a referendum in Denmark in July 1992 made the continuance of the present set of parities seem rather less likely. The initial problems occurred in September 1992, when Finland and Sweden faced problems as a result of their financial crises, which had erupted a year earlier. Although not members of the EU, they had both pegged their currencies to the ECU under the same terms as the member states, although of course without any commitment to reciprocal intervention. Finland floated, allowing a devaluation, but Sweden fought off the challenge by applying exceptionally high interest rates (the respite was only temporary and Sweden too floated after a second attack in November). Attention then turned to Italy, which was forced to devalue its parity. However, within a few days that new parity also looked unsustainable and the attacks were extended to the UK and Spain. On Black Wednesday, 16 September, the losses proved too great and the UK left the ERM, followed by Italy, but Spain, although devaluing, continued in the ERM. On 20 September France narrowly voted in favour of ratifying the Maastricht Treaty in a referendum. Speculation beforehand that the result might be rejection is likely to have contributed to the exchange rate pressures. But although France was itself subject to pressure it survived. Ireland, Portugal and Spain introduced exchange controls.

The exchange rate speculation continued and in mid-November Sweden gave up the struggle and floated. Spain and Portugal devalued three days later. Ireland devalued in January 1993, Spain and Portugal devalued again in May. However, it was clear that the pressure on all the remaining currencies except the guilder was likely to force further devaluations, and on 1 August the permitted bands of fluctuation were extended from $\pm 2.25\%$ to $\pm 15\%$, which was sufficient to cope with the misalignment and more importantly allowed exchange rates to move far enough to choke off market pressure. (The Netherlands was able to continue with the narrower band.) While technically the ERM had survived, in practice floating had been required to survive the crisis. Fixing of exchange rates only came back as the member states moved to monetary union itself under the terms of the Maastricht Treaty.

The Road to the Euro Area

The increasing success of the EMS in the 1980s brought the ideas of moving towards monetary union onto the agenda again. At the beginning of

his first Presidency of the Commission, Jacques Delors had canvassed a number of ideas for greater progress on European integration. The idea of completing the internal (or single) market came first, but monetary union was second. However, closer integration of product markets, which included financial services and labour markets as part of the single market programme, which was incorporated into the Single European Act in 1987, itself helped pave the way for monetary integration. The closer integration increased the extent to which the member states were likely to move together in economic fluctuations and increased the ability to respond to shocks through increased flexibility. The decision to try to move forward was made at the Hanover Council in June 1988. As in 1970 this also took the form of a Committee, to be chaired by Jacques Delors. However, the composition of the committee was completely different. It was composed of the central bank governors from each of the member states and two experts. This meant that it was dominated by the practical considerations of how to move to a single currency from the monetary perspective.

The resulting proposals had significant differences from their 1970 counterparts. In particular, they focused on creating the appropriate EU level institutions to implement such a currency. However, the framework was still one of economic and monetary integration, although the economic side did not involve matching EU level institutions. It was also to be achieved in three stages, echoing the Werner Report. The first stage was to concentrate on fiscal consolidation, greater convergence of macroeconomic policy and performance through closer coordination, completion of the single market, greater financial integration and coordination of monetary policies. The second stage, which would require a treaty change, would set up a European System of Central Banks (ESCB) and involve national monetary policies being executed with EC-level objectives in mind and harmonization of the tools of monetary policy. In the third stage, exchange rates would be irrevocably fixed and the ESCB would assume responsibility for monetary and exchange rate policy, with a pooling of reserves.

It was agreed to proceed with these ideas at the Madrid summit in June 1989 and commence Stage 1 on 1 July 1990 after the details had been sorted out by an intergovernmental conference. This conference produced the Treaty on European Union, which was approved at the Maastricht Council in December 1991.

The Treaty clarified two main issues: first it announced the creation of the European Monetary Institute, which was to come into being at the start of the second stage to prepare all the instruments and procedures for the single monetary policy to be followed by the European Central Bank as the EC level institution to implement policy in Stage 3; second it set out the timetable and a set of criteria for joining Stage 3. Stage 2 was to start in 1994 and Stage in 1997 if seven or more of the member states met the convergence criteria, or failing that in 1999 with as many states as met the criteria.

There were five criteria

- *Price stability*: 'a price performance that is sustainable and an average rate of inflation, observed over a period of one year before the examination that does not exceed by more than 1.5 percentage points that of, at most, the three best performing member states'.
- *Interest rates*: 'over a period of one year before the examination. . . an average nominal long-term interest rate that does not exceed by more than two percentage points that of, at most, the three best performing member states in terms of price stability'.
- *Budget deficits*: the member state must not have an 'excessive deficit', which a protocol attached to the treaty defines as 3% of GDP but Council can override this if 'either the ratio has declined substantially and continuously and reached a level that comes close to the reference value; or. . . the excess over the reference value is only exceptional and temporary'.
- *Public debt*: the ratio of government debt should not exceed 60% of GDP 'unless the ratio is sufficiently diminishing and approaching the reference value at a satisfactory pace'. The proviso was to be decided by the Council (using qualified majority voting).

(It was also noted that the Council should take a medium term view, so just passing the reference value on the assessment date might not be enough.)

- *Currency stability*: 'respected the normal fluctuation margin provided for by the exchange rate mechanism... without severe tension for at least two years before the examination'.

These constitute a rather narrow 'monetary' view of what constitutes adequate convergence to be able to join a monetary union. This is clearly distinct from an 'economic' view, which would require real convergence in a number of key respects and the ability to respond flexibly to future shocks (described as asymmetric or idiosyncratic shocks that would affect that member state but not the EU as a whole) without the ability to use monetary policy. The real convergence would imply economic structures, policies and income levels. If these three were too different then a common policy would impose undue strains on the member state from common shocks that affected the whole monetary union. This focus on the monetary characteristics no doubt reflects the composition of the Delors Committee, but it also helps explain subsequent difficulties with the union.

Even within the monetary framework, there is no direct explanation of the fiscal criteria. A 60% debt to GDP ratio was around the average prevailing in the EU at the time and, with a 3% growth rate, a deficit of 3% a year would not worsen the position (after making a fairly sweeping assumption about the relationship between inflation rates and interest rates). Hence setting this as a minimum requirement would tend to imply an improving debt position. Furthermore, by restricting convergence to inflation rather than the price level left open the problem that, as countries move closer to a genuine single market, price levels can be expected to converge, which entails that subsequent inflation rates will be different until that process is complete.

The process of negotiating and ratifying the treaty was not straightforward. The UK insisted on an opt-out from the requirement to join monetary union, and a similar opt-out was accorded to Denmark after the rejection of the treaty in a referendum (a second referendum narrowly approved the revised proposals in May 1993). Thus although the treaty was agreed in December 1991 in Maastricht and signed in February 1992 in Limburg, it did not ultimately come into force until November 1993.

The process of convergence proved difficult, and at the first date of assessment in June 1996, for commencement in 1997, only Luxembourg qualified so there was no attempt at a formal examination. The EU had in the meantime increased its membership by three in 1995 with the accession of Austria, Finland and Sweden. By June 1998 the picture was very different. Finland, France and Luxembourg met all the criteria on a strict interpretation and Austria, Germany, Ireland, the Netherlands, Portugal and Spain were close to meeting the fiscal criteria. Belgium and Italy, with debt ratios of 122.2% and 121.6% respectively, were nowhere near, but nevertheless were admitted.

Greece had inflation 1.5% above the convergence criterion (and a debt ratio of 108.7% and a deficit of 4%), Denmark and the UK exercised their opt-outs (although both could have converged) and Sweden, having had membership of EMU rejected in a referendum, remained outside the ERM, thereby technically failing to qualify.

The Maastricht Treaty concentrated on the monetary side of EMU. Unlike the Werner recommendations, it did not agree for there to be an organization for the coordination or management of fiscal or other macroeconomic policies nor their interrelationship with monetary policy. Instead, it set out a general requirement for coordination of economic policy: 'Member States shall regard their economic policies as a matter of common concern and shall co-ordinate them within the Council'. Over the ensuing years the process of macroeconomic coordination was slowly developed into a comprehensive scheme by a series of 'processes' whose individual names reflect the location of the Council meetings at which they were agreed: the Luxembourg process coordinating employment policies (1997); the Cardiff process coordinating structural policies (1998); and the Cologne process (1999) coordinating macroeconomic policies. These all operate under the

framework of the Broad Economic Policy Guidelines, which were originally discussed annually but have subsequently had a three-year horizon. The important feature of these processes is that they do not compel action but draw up common objectives, areas of focus and principles for action, in what is known as the Open Method of Coordination. The Commission then monitors progress against the targets.

While the member states could manage to 'coordinate' their fiscal policies with the single monetary policy if the latter had clear objectives and a transparent and clearly articulated strategy for implementation, it was not felt possible to proceed simply through the coordination process, as countries might otherwise run profligate fiscal policies that would damage the creditworthiness of EMU as a whole. At the same time as the idea of the processes was launched at the Dublin Council in 1996, it was agreed to seek an approach to fiscal stability which in effect would ensure that the fiscal criteria laid down for entry to Stage 3 of EMU were perpetuated during its operation. This latter approach was agreed as the Stability and Growth Pact in 1997. The Pact had two elements to it, which have been labelled 'preventative and corrective'. The preventative part involves the setting of longer term objectives for prudent fiscal policy and an annual process of surveillance by the Commission on progress, including the shorter term progress across the cycle. Thus the longer term aim is to keep the debt position steadily improving by ensuring that budgets are 'close to balance or in surplus' across the economic cycle and in the shorter term ensuring that at no stage does the budget deficit fall below 3% unless the country concerned is under substantial economic pressure.

The corrective part of the Pact involves the rules for avoiding and correcting any such 'excessive deficits' and is therefore labelled the Excessive Deficit Procedure. In the annual cycle of surveillance the Commission can opine that a country is likely to encounter an excessive deficit and then the Council of Economic and Finance Ministers (ECOFIN) can recommend that remedial action should be taken. If that action is not taken or is insufficiently applied, ECOFIN can in theory require an interest-free deposit of up to 0.5% of GDP and could convert this into a fine. This, however, has never been applied in practice and the lenient treatment of France and Germany when they got into difficulty in 2003 and the subsequent revision of the Pact in 2005 to make the criteria for excessive deficits softer could be taken to suggest that such sanctions are unlikely in future.

The Development of Monetary Union

An analysis of how Stage 3 of EMU has evolved since its inception at the beginning of 1999 lies beyond the scope of this article, but the experience of the period up to 2011 provides some insights into the structure and development of EMU up to that point. Three items stand out. The first is that in technical terms the design of the monetary side of EMU has been shown to be exemplary. There were no technical slip-ups or instability in financial markets and both the single monetary policy and the currency came into operation exactly as planned. Thus the framework set out by the committee of central bankers and the staged implementation through the EMI provided a workable template that others could build on. Having the institutional arrangement at the EU level was essential. The second issue that stands out is that the economic side of EMU has not proven particularly successful. There is no matching institution, as envisaged in the Werner framework. While fiscal behaviour since the mid-1990s has been a great improvement on that before, Greece and (to a lesser extent) Portugal have been unable to impose the fiscal stability desired and an emergency lending programme has been required in concert with the IMF to enable them to meet their debt obligations. (Ireland has also had debt problems, but to a major extent these are due to bad banking supervision and crisis management and only partly to an optimistic fiscal policy relying on continuing rapid growth.)

The third insight is that a focus on a narrow view of monetary convergence as a precondition for a successful economic and monetary union rather than a focus on the economic optimum currency area criteria has provided difficulties.

Countries faced by different shocks in the global financial crisis have had difficulty adjusting as they no longer have the exchange rate as an adjustment mechanism. Countries joining with lower than average real income per head and lower price levels have found that while they experience faster growth, they also face faster inflation, as monetary policy focuses only on the rate of inflation for the EU as a whole. Since the MacDougall Report in 1977, the EU has shown little enthusiasm for fiscal equalization across countries and has not pursued anything like the scale of fiscal transfers observed in other large diverse countries with a monetary union, such as Australia, Canada, the USA and even Germany. Thus this route to adjustment, widely used elsewhere, has also not been available. The criteria for membership and their interpretation in 1998 reflected the political pressure for monetary union rather than simply an economic assessment. The changes to the proposed process of economic and monetary integration after the Werner Committee report in 1970 reflected the wish by the member states for more economic policy autonomy while pursuing tight exchange rate stability.

However, no satisfactory test of the success of EMU is possible, as it requires the ability to simulate a credible alternative, which would be a purely hypothetical exercise. Thus one can neither estimate with any reliability how the chosen form of EMU has fared by comparison with a Werner-style arrangement, perhaps with harsher criteria and hence fewer members or with a slower process that required a reasonable degree of real convergence and adherence to the optimum currency criteria. While the Werner vision may have taken three decades, and not one, to implement, EMU has been able to progress far faster and further than would have been thought likely in say 1985.

See Also

- ▶ European Central Bank
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Cohesion Policy
- ▶ European Monetary Union

Bibliography

Ardy, B., I. Begg, D. Hodson, I. Maher, and D.G. Mayes. 2006. *Adjusting to EMU*. Basingstoke: Palgrave Macmillan.

Artis, M., and M. Buti. 2000. 'Close to balance or in surplus': A policy-maker's guide the implementation of the stability and growth pact. *Journal of Common Market Studies* 38: 563–591.

Begg, I., D. Hodson, and I. Maher. 2003. Economic policy co-ordination in the European Union. *National Institute Economic Review* 193: 66–77.

Commission of the European Communities. 1969. *Commission Memorandum to the Council on the co-ordination of economic policies and monetary co-operation within the Community,* Barre Report. Brussels: Commission of the European Communities.

Commission of the European Communities. 1970. Report to the Council and the Commission on the realization by stages of economic and monetary union in the Community. *EU Bulletin*, supplement 11, the Werner Report. Luxembourg: Commission of the European Communities.

Commission of the European Communities. 1977. *Report of the study group on the role of public finance in European integration,* the MacDougall Report. Brussels: Commission of the European Communities.

Commission of the European Communities. 1999. Convergence report 1998. *European Economy* 65: 23.

De Grauwe, P. 2009. *Economics of monetary union*, 8th ed. Oxford: Oxford University Press.

Dyson, K., and K. Featherstone. 1999. *The road to Maastricht: Negotiating economic and monetary union*. Oxford: Oxford University Press.

Emerson, M., D. Gros, A. Italianer, and H. Reichenbach. 1991. *One market one money: An evaluation of the potential benefits of forming an economic and monetary union*. Oxford: Oxford University Press.

Frankel, J., and A. Rose. 1998. The endogeneity of the optimum currency area criteria. *Economic Journal* 108: 1009–1025.

Holzmann, H. (ed.). 1996. *Maastricht: Monetary constitution without a fiscal constitution*. Baden-Baden: Nomos Verlag.

Issing, O. 2002. On macroeconomic policy coordination. *Journal of Common Market Studies* 40: 345–358.

Kenen, P. 1969. A theory of optimum currency areas: An eclectic view. In *Monetary problems of the international economy*, ed. R.A. Mundell and A.K. Swoboda. Cambridge, MA: MIT Press.

Ludlow, P. 1982. *The making of the European Monetary System: A case study of the politics of the European Community.* London: Butterworths.

Mundell, R.A. 1973. A plan for a European currency. In *The economics of common currencies*, ed. H.G. Johnson and A.K. Swoboda. London: Allen and Unwin.

E

European Monetary Union

Paul De Grauwe

Abstract

The introduction of the euro in 1999 is without any doubt one of the great achievements in the European integration process. In one bold stroke, national monetary sovereignty was abolished and transferred to a new European institution, the European Central Bank, that from then on became the guardian of the new currency.

Until the eruption of the sovereign debt crisis there was a general perception that the euro zone was a great success. In 2008 the European Commission issued a report (euro@10; European Commission. *Europe@10. Successes and challenges after ten years of economic and monetary unions*. Brussels, 2008) that was unqualified in its praise about the achievements of the euro zone. Then came the sovereign debt crisis that has led many observers to reevaluate European Monetary Union (EMU). This article discusses its successes and failures, analyzes the fragility of EMU, and identifies two sources of this fragility. Finally, it discusses governance issues and the nature of the political institutions that will be necessary to sustain the European Monetary Union.

Keywords

European Central Bank (ECB); European Financial Stability Facility (EFSF); European Stability Mechanism (ESM); Euro; Housing market; Sovereign debt crisis; Stability and Growth Pact

JEL Classifications

E42; E5; E50; E52; G1

Successes of EMU

The very fact that European countries managed to move into a monetary union using a peaceful process can be seen as an important historical achievement. Most of the monetary unions in history came about as a result of military conquest or forceful political unification. This was not the case in Europe during the 1990s when monetary integration process was set in motion. A massive transfer of monetary sovereignty was successfully organized, leading to the establishment of the European Central Bank, which was given the task to manage a common currency: the euro.

Up until the eruption of the sovereign debt crisis, the successes appeared overwhelmingly strong. The benefits of a common currency, which were analyzed in the theory of optimal currency areas, could not easily be disputed. The use of one currency in the euro zone eliminated the transaction costs that existed prior to union and that arose from the fact that in order to make trade possible between two member countries one national currency had to be exchanged for another. The EC Commission (1990) estimated that the elimination of these transaction costs amounted to approximately half a per cent of GDP.

There can be equally little doubt that the elimination of exchange risk within the euro zone helped to boost internal trade and capital mobility. A lot of research has been done to measure the effect of a monetary union on trade between the members of the union. Following Andy Rose's ground-breaking research (Rose 2000), which demonstrated strong positive effects of monetary unions on trade flows, subsequent econometric research has confirmed that monetary unions in general, and EMU in particular, indeed lead to significant increases in trade. However, the Rose's spectacular results were not replicated in subsequent work (see Baldwin 2006; Berger and Nitsch 2008). The consensus today seems to be that EMU may have added approximately 20 per cent of extra trade within the union; a significant increase that certainly should be added to the successes of the Union.

The institutional setup of the euro zone also contributed to the successes of EMU. By giving the ECB a strong mandate to maintain price stability, and by enshrining the political independence of the ECB in the EU Treaties, the ECB quickly gained credibility as a tough inflation fighter. There can be little doubt that the ECB was very

successful in keeping inflation low. From 1999 to 2010 the average inflation rate in the euro zone was 2.2 per cent – not much above the target of 2 per cent that the ECB had set itself as an objective, and certainly lower than the rate of inflation its members experienced during the post-war period until the start of the euro zone. As a result, the euro zone became a centre of price stability.

Behind this apparent success, however, there are deep structural weaknesses that have appeared with full force since the eruption of the sovereign debt crisis. These are discussed in the next section.

Failures of EMU

Two structural weaknesses lie at the heart of the sovereign debt crisis that began in 2009. The first one arises from asymmetric shocks and the absence of flexible adjustment mechanisms. This is the feature that has been stressed by the Optimal Currency Area (OCA) theory as generating costs of a monetary union (well-known contributions to this theory are Mundell (1961) and McKinnon (1963). For surveys see Ishiyama (1975), De Grauwe (1992) and Baldwin and Wyplosz (2006). See Eichengreen (1990) and Bayoumi and Eichengreen (1996) for empirical implementations). The second structural weakness arises from the fact that the euro is a currency without a country.

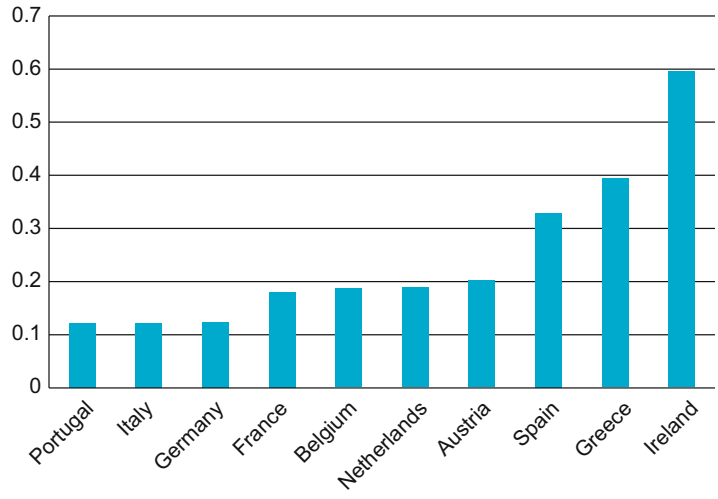
Failing Adjustment to Asymmetric Shocks

In EMU, monetary policies are centralized, and therefore cease to be a source of asymmetric shocks. The member countries of EMU, however, continue to exercise considerable sovereignty in several economic areas. The most important one is in the budgetary field. The spending and taxing powers in EMU continue to be vested in the hands of national authorities. Today, in most euro zone countries, spending and taxation by the national authorities amount to close to 50 per cent of GDP. The spending and taxing powers of the European authorities represent barely 1 per cent of GDP. This situation has not changed since the start of monetary union in 1999. By changing taxes and spending, the authorities of an individual country can create large asymmetric shocks. By their very nature these shocks are well contained within national borders. For example, when the authorities of a country increase taxes on wage income, this only affects labour in that country and will influence spending and wage levels in that country. As a result, asymmetric shocks are created that lead to necessary adjustments later.

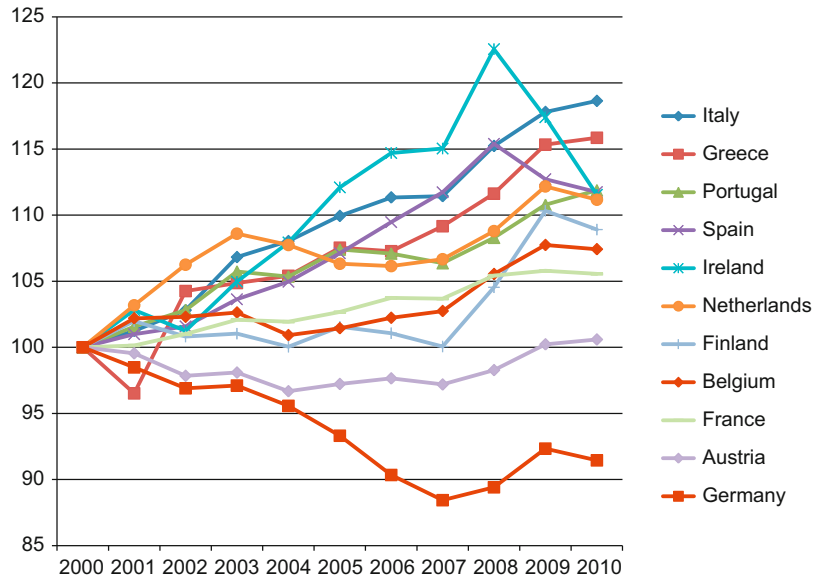
There are other aspects of the existence of nation states that can be a source of asymmetric disturbances. Many economic institutions are national. Wage bargaining systems, for example, differ widely between countries, creating the possibility of asymmetric disturbances. In addition, differences in legal systems and customs generate significant differences in the workings of financial markets. For example, regulations about the conditions under which mortgages are granted by banks differ from one country to the other in the euro zone. These differences can lead to very divergent movements of housing prices in member countries. There are many more such examples of these asymmetric disturbances.

The effect of these national idiosyncrasies can be that countries experience very different economic conditions. This is illustrated in Fig. 1, which shows the cumulative growth rates of GDP prior to the financial crisis in the euro zone. We observe that indeed the differences in national growth rates in the euro zone were substantial. Some countries experienced booming economic conditions during 1999–2007 (Spain, Greece, Ireland); others experienced very slow growth (Portugal, Italy, Germany). There can be little doubt that part of these differences is attributable to the different national economic policies and institutions.

These diverging trends in economic activity can contribute to another important phenomenon: the emergence of large divergences in competitive positions of members of a monetary union. This is exactly what happened in the euro zone. Figure 2 illustrates this by presenting the trends in the relative unit labour costs in the euro zone during 2000–2010. The relative unit labour cost is defined as the unit labour cost of one country (say Germany) relative to the average unit labour



European Monetary Union, Fig. 1 Cumulative growth of GDP 1999–2007 (Source: European Commission; AMECO databank)



European Monetary Union, Fig. 2 Relative unit labour costs for euro zone countries (2000 = 100). (Source: European Commission, AMECO databank)

costs in the other member countries of the euro zone. When the relative unit labour cost declines, as in the case of Germany, one can say that Germany has improved its competitive position during the period 2000–2010. Conversely when the relative unit labour costs increase, as in the cases of Ireland, Italy, and Greece, among others, these countries lost competitiveness during 2000–2010.

Note that the unit labour costs are defined as the labour costs corrected for labour productivity. The unit labour cost is defined as: ULC = $W/(Q/L)$, where W is the wage rate, Q is the value of output and L is the amount of labour used in production. Q/L is the average labour productivity. Note that the formula can also be rewritten as follows: ULC = WL/Q. This expression makes clear what unit labour costs means: it is the wage cost embedded in one euro (dollar) of output. It follows that unit labour costs can increase for two reasons. They increase when wages increase or when labour productivity declines (or, in relative terms fails to keep pace with competitors). Thus it appears that countries like Italy, Spain, Portugal and Ireland have lost significant competitiveness

since 2000 because wages in these countries increased faster than labour productivity. This leads to serious adjustment problems. These countries have to reduce their wage levels relative to the other countries of the euro zone (if they cannot raise productivity). This is likely to be a slow and painful process, mainly because other adjustment mechanisms such as labour mobility and wage flexibility function poorly in most of these countries (see Krugman (1993) and De Grauwe and Vanhaverbeke (1990)). In the past, when these countries were not in a monetary union, they would have been able to devalue their currencies, thereby making it possible to restore their competitiveness in a single stroke, albeit only by risking higher inflation.

The Euro: A Currency Without a Country

The second major structural weakness of EMU is that it created a currency without a country, i.e. without a government with the full powers of a government to back up the value of the currency. This feature has sometimes been hailed as revolutionary. It turns out, however, that it is a profound structural weakness that creates great fragility in the euro zone and lies at the heart of the sovereign debt crisis. Let us analyze why this is so.

When entering monetary union, member countries lose their capacity to issue debt in a currency over which they have full control. As a result, a loss of confidence of investors can, in a self-fulfilling way, drive the country into default (Kopf 2011). The reason for this can be described as follows. Suppose that investors fear a default by, let's say, the Spanish government. They sell Spanish government bonds, thereby raising the interest rate. The investors who have acquired euros are likely to decide to invest these euros elsewhere, perhaps in German government bonds. As a result, the euros leave the Spanish banking system. Thus the total amount of liquidity (money supply) in Spain shrinks. The Spanish government experiences a liquidity crisis, i.e. it cannot obtain funds to roll over its debt at reasonable interest rates. In addition, the Spanish government cannot force the Bank of Spain to buy government debt. The ECB can provide all the liquidity in the world, but the Spanish government does not control that institution.

This is not the case for countries capable of issuing debt in their own currency. Let's trace what would happen if investors were to fear that the UK government might be defaulting on its debt. In that case, they would sell their UK government bonds, driving up the interest rate. After selling these bonds, these investors would have pounds that they would most probably want to get rid of by selling them in the foreign exchange market. The price of sterling would drop until somebody else would be willing to buy them. The effect of this mechanism is that the pounds would remain bottled up in the UK money market to be invested in UK assets. Put differently, the UK money stock would remain unchanged. Part of that stock of money would probably be reinvested in UK government securities. But even if that were not the case, and the UK government was unable to find the funds to roll over its debt at reasonable interest rates, it would certainly force the Bank of England to buy up the government securities. Thus the UK government is assured of the liquidity needed to fund its debt. This means that investors cannot precipitate a liquidity crisis in the UK that could, ultimately, push the UK government into default. There is a superior force of last resort, namely the Bank of England.

This different mechanism explains why the Spanish government has been obliged to pay up to 200 basis points more on its ten-year bonds than the UK government since 2010, despite the fact that its government debt and deficit were significantly lower than the UK ones.

Because of the liquidity flows triggered by changing market sentiments, member countries of a monetary union become vulnerable to these market sentiments. These can lead to 'sudden stops' in the funding of government debt (Calvo 1988), setting in motion a devilish interaction between liquidity and solvency crises. This is because the liquidity crisis raises the interest rate, which in turn leads to a solvency crisis. This problem is not unique to members of a monetary union. It has been found to be very important in emerging economies that cannot issue debt in their own currencies

(see Eichengreen et al. (2005), in which these problems are analyzed in great detail).

There are further important implications of the increased vulnerability of member countries of a monetary union. In De Grauwe (2011) these implications are developed in greater detail; see also Wolf 2011). One of these implications is that member countries of a monetary union lose much of their capacity to apply countercyclical budgetary policies. When, during a recession, budget deficits increase, this risks creating a loss of confidence of investors in the capacity of the sovereign to service the debt. This has the effect of raising the interest rate, making the recession worse and leading to even higher budget deficits. As a result, countries in a monetary union can be forced into a bad equilibrium, characterized by deflation, high interest rates, high budget deficits and a banking crisis (see De Grauwe (2011) for a more formal analysis).

What Kind of Governance?

The previous discussion points towards the existence of a coordination failure in EMU. Financial markets can drive member countries into a bad equilibrium that is the result of a self-fulfilling mechanism. This coordination failure can in principle be solved by collective action aimed at steering countries towards a good equilibrium, but as the difficulties in dealing with the Greek crises of 2010 and 2011 showed, taking collective action is always politically difficult.

In addition to this coordination failure, there is another important feature of the euro zone that requires collective action. This is that the euro zone creates externalities, especially through contagion. When one country is pushed into a bad equilibrium, this affects all the other countries, mainly because of the intense degree of financial integration. As a result, a default risk in one country can lead to a default risk of sovereigns and banks in other countries. As with all externalities, government action must be resolute in internalizing these.

Collective action and internalization can be pursued at two levels. One is at the level of the central banks; the other at the level of government budgets.

Liquidity crises are avoided in standalone countries that issue debt in their own currencies, mainly because the central bank can be forced to provide all the necessary liquidity to the sovereign. This outcome can also be achieved in a monetary union if the common central bank is willing to buy the different sovereigns' debt. In fact, this is what happened in the euro zone during the debt crisis. The ECB bought government bonds of distressed member countries, either directly or indirectly by the fact that it accepted these bonds as collateral in its support of the banks from the same distressed countries. In doing so, the ECB rechannelled liquidity to countries hit by a liquidity crisis, and prevented the centrifugal forces created by financial markets from breaking up the euro zone. It was the right policy for a central bank whose *raison d'être* it is to preserve the monetary union. Yet the ECB has been severely criticized for saving the euro zone this way. The main reason for this criticism is that these liquidity provisions have potential fiscal policy consequences. For example, when the ECB buys Greek and Portuguese government bonds in order to rechannel liquidity to Greece and Portugal, it exposes itself to the risk of future losses. In doing so it commits euro zone taxpayers to foot the bill in the future, without having asked their permission. This criticism has been powerful enough to convince the ECB that it should not be involved in such liquidity operations, and that liquidity support must instead be done by other institutions, in particular a European Monetary Fund.

An important step was taken in May 2010 when the European Financial Stability Facility (EFSF) was instituted. The latter will be transformed into a permanent fund, the European Stabilization Mechanism (ESM), which will obtain funding from the participating countries and will provide loans to countries in need of liquidity assistance. This makes it possible to make the fiscal commitments of each country explicit. Thus a European Monetary Fund will be in existence, as was first proposed by Gros and Mayer (2010).

Although an important step forward, the EFSF, as well as the future ESM, suffer from problems

that undermine their effectiveness. The most important one is that neither will be an autonomous institution in the way that the IMF is. Each country keeps its veto power for every new financial assistance program. This feature risks making these institutions less than fully effective. As a result, the credibility of the institution will be undermined, as nobody knows whether and under what conditions the EFSF (ESM) will be willing to provide credit. The only way to solve this problem is to transform the EFSF (ESM) into a true monetary fund in which decisions are taken by qualified majority, as is the case in other European institutions (e.g. the Council of Ministers). This, of course, implies that there will be a willingness to transfer sovereignty to the monetary fund.

Collective action and internalization can also be taken at the budgetary level. Ideally, a budgetary union is the instrument of collective action and internalization. By consolidating (centralizing) national government budgets into one central budget a mechanism of automatic transfers can be organized. Such a mechanism works as an insurance mechanism transferring resources to the country hit by a negative economic shock. In addition, such a consolidation creates a common fiscal authority that can issue debt in a currency under the control of that authority. In so doing, it protects the member states from being forced into default by financial markets. It also protects the monetary union from the centrifugal forces that financial markets can exert on the union. The need to create a budgetary union together with a monetary union has long been recognized by economists (McDougall Report, 1977; Sachs and Sala-i-Martin 1989; Mélitz and Vori 1993; Von Hagen 1996). However, monetary union was started in Europe without such a budgetary union, creating the fragility discussed earlier.

While a full budgetary union is not a realistic prospect in the euro zone in the foreseeable future, smaller steps could be taken, signalling a desire to move towards budgetary union in the future. One such step consists in the joint issue of Eurobonds. By jointly issuing Eurobonds, the participating countries become jointly liable for the debt they have issued together. This is a very visible and constraining commitment that may help to convince financial markets that member countries are serious about the future of the euro (Verhofstadt 2009; Juncker and Tremonti 2010). In addition, by pooling the issue of government bonds, the member countries protect themselves against the destabilizing liquidity crises that arise from their inability to control the currency in which their debt is issued. A common bond issue does not suffer from this problem. Several concrete proposals have been formulated by Bruegel (Delpla and von Weizsäcker 2010; De Grauwe and Moesen 2009). These also discuss the inevitable moral hazard issues that arise with the implementation of the common Eurobond issues (see also Issing (2009) and Gros (2011).

It should be noted that if successful, such a common Eurobond issue would create a large new government bond market with a lot of liquidity. This in turn would attract outside investors, making the euro a reserve currency. It has been estimated that the combined liquidity and reserve currency premium enjoyed by the US dollar amounts to approximately 50 basis points (Gourinchas and Rey 2007). A similar premium could be enjoyed by the euro. This would make it possible for the eurozone countries to lower the average cost of borrowing, very much like the USA has been able to do.

Another important step in the process towards political union is to set some constraints on the national economic policies of the member states of the euro zone. As argued earlier, the fact that, while monetary policy is fully centralized, the other instruments of economic policies have remained firmly in the hands of national governments is a serious design failure of the euro zone. Ideally, countries should hand over sovereignty over the use of these instruments to European institutions. However, the willingness to take such a drastic step towards political union is completely absent. Here also small steps should be taken.

Some progress has been achieved in setting up new rules of economic governance in the euro zone. A so-called 'six pack' of measures strengthening control on budgetary policies and coordinating macroeconomic policies is very likely to be adopted. These measures include a tightening of the Stability and Growth Pact, including a

stronger sanctioning procedure; the 'European Semester', which requires national governments to present their annual budgets to the European Commission prior to their approval in national parliaments; and the monitoring of a number of macroeconomic variables (current account balances, competitiveness measures, house prices and bank credit) aimed at detecting and redressing national macroeconomic imbalances. Failure to take action to eliminate these imbalances could trigger a sanctioning mechanism very much in the spirit of the sanctioning mechanism of the Stability and Growth Pact.

The proposals for reforming the governance or the euro zone that have been discussed in this section all require a far-reaching degree of political union. Economists have stressed that such a political union will be necessary to sustain the monetary union in the long run (EC Commission 1977; De Grauwe 1992). It is clear, however, that there is little willingness in Europe today to increase the degree of political union substantially, although the decisions taken by the heads of state and government of the euro area at their meeting on 21 July 2011 represented a shift in this direction. This unwillingness to go significantly further in the direction of more political union will continue to make the euro zone a fragile construction.

Conclusion

Any monetary union creates benefits and costs. This is also the case in the European Monetary Union. The benefits that were created in the euro zone are significant. They arise from the fact that the elimination of national currencies reduces transaction costs and eliminates the uncertainty produced by exchange rate volatility.

The costs, however, are also substantial. They arise from the fact that member countries of a monetary union lose an instrument of economic policy that can help countries to adjust to asymmetric shocks. In addition, countries that join a monetary union lose their capacity to issue debt in a currency over which they have full control. As a result, a loss of confidence of investors can, in a self-fulfilling way, drive the country towards default. This is not so for countries capable of issuing debt in their own currency. In these countries the central bank can always provide liquidity to the sovereign to avoid default. This may lead to future inflation, but it shields the sovereign from a default forced by the market.

Thus member countries of a monetary union become very vulnerable to changing market sentiments. The latter can lead to 'sudden stops' in the funding of the government debt, setting in motion a devilish interaction between liquidity and solvency crises. This feature of a monetary union creates great fragility in sovereign debt markets in a monetary union. This fragility can only be overcome by collective action.

A monetary union can only function if there is a collective mechanism of mutual support and control. Such a collective mechanism exists in a political union. In the absence of a political union, the member countries of the euro zone are condemned to fill in the necessary pieces of such a collective mechanism. The debt crisis has made it possible to fill in a few of these pieces. What has been achieved, however, is still far from sufficient to guarantee the survival of the euro zone.

See Also

- ▶ Euro Zone Crisis 2010
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Central Bank
- ▶ European Monetary Integration
- ▶ European Union Budget
- ▶ Subprime Mortgage Crisis

Acknowledgments I am grateful to Iain Begg and Alison Howson for comments and suggestions.

Bibliography

Baldwin, R. 2006. The euro's trade effects. In *ECB working paper, no. 594*. Frankfurt: European Central Bank.

Baldwin, R., and C. Wyplosz. 2006. *The economics of European integration*. 2nd ed. London: McGraw-Hill.

Bayoumi, T. and Eichengreen, B. 1996. Operationalizing the theory of optimum currency areas. *CEPR discussion paper,* no. 1484.

Berger, H. and Nitsch, V. 2008. Zooming out: The trade effect of the euro in historical perspective. *Journal of International Money and Finance* 27(8): 1244–1260.
Calvo, G. 1988. Servicing the public debt: The role of expectations. *American Economic Review* 78(4): 647–661.
De Grauwe, P. 1992. *The economics of monetary integration*. 1st ed. Oxford: Oxford University Press.
De Grauwe, P. 2011. The governance of a fragile eurozone. http://www.econ.kuleuven.be/ew/academic/intecon/Degrauwe/PDG-papers/Discussion_papers/ Governance-fragile-eurozone_s.pdf
De Grauwe, P. and Moesen, W. 2009. Gains for all: A proposal for a common Eurobond. *Intereconomics* 132–141.
Delpla, J. and von Weizsäcker, J. 2010. The blue bond proposal. *Bruegel policy brief*.
EC Commission (1977). Report of the study group on the role of public finance in European integration (MacDougall Report), Brussels.
EC Commission. 1990. One market, one money. *European Economy* 44.
Eichengreen, B. 1990. Is Europe an optimum currency area? *CEPR discussion paper*, no. 478.
Eichengreen, B., R. Hausmann, and U. Panizza. 2005. The pain of original sin. In *Other people's money: Debt denomination and financial instability in emerging market economies*, ed. B. Eichengreen and R. Hausmann. Chicago: Chicago University Press.
European Commission. 2008. *Europe@10. Successes and challenges after ten years of economic and monetary unions*. Brussels.
European Commission. 2010. *A structured framework to prevent and correct macroeconomic imbalances: Operationalizing the alert mechanism*. Brussels.
Gourinchas, P.-O., and H. Rey. 2007. From world banker to world venture capitalist: The US external adjustment and the exorbitant privilege. In *G7 current account imbalances: Sustainability and adjustment*, ed. R. Clarida. Chicago: University of Chicago Press.
Gros, D. 2010. The seniority conundrum: Bail out countries but bail in private short-term creditors. *CEPS Commentary* 10.
Gros, D. and Mayer, T. 2010. Towards a European monetary fund. *CEPS Policy Brief.* http://www.ceps.eu/book/towards-european-monetary-fund.
Ishiyama, Y. 1975. The theory of optimum currency areas: a survey. *IMF Staff Papers* 22: 344–383.
Issing, O. 2009. Why a common eurozone bond isn't such a good idea. *Europe's World* 677–679.
Juncker, J.-C. and Tremonti, G. 2010. E-bonds would end the crisis. *Financial Times* 5.
Kopf, C. 2011. Restoring financial stability in the euro area. *CEPS Policy Briefs* 15.
Krugman, P. 1993. Lessons of Massachusetts for EMU. In *Adjustment and growth in the European monetary union*, ed. F. Torres and F. Giavazzi. London/ Cambridge: CEPR/Cambridge University Press.
McKinnon, R. 1963. Optimum currency areas. *American Economic Review* 53: 717–725.
Mélitz, J., and S. Vori. 1993. National insurance against unevenly distributed shocks in a European Monetary Union. *Recherches Economiques de Louvain* 59: 1–2.
Mundell, R. 1961. A theory of optimal currency areas. *American Economic Review* 51.
Rose, A.K. 2000. One money, one market: Estimating the effect of common currencies on trade. *Economic Policy* 30: 9–45.
Sachs, J. and Sala-i-Martin, X. 1989. Federal fiscal policy and optimum currency areas. *Harvard University Working Paper.* Cambridge, MA.
Verhofstadt, G. 2009. *De Weg uit de Crisis. Hoe Europa de wereld kan redden*. Amsterdam: De Bezige Bij.
Von Hagen, J. 1996. *Währungsunion, Fiskalunion, Politische Union.* Mimeo, Universität Mannheim.
Wolf, M. 2011. Managing the eurozone's fragility. *Financial Times* 4. http://www.ft.com/cms/s/0/f30b14f8-75ab-11e0-80d5-00144feabdc0.html#axzz1LMlUathU.

European Unemployment Insurance

Sebastian Dullien

Abstract

'European unemployment insurance' is one of many proposals that aim to create a countercyclical automatic fiscal stabiliser for the euro area. References to such systems were made in the 1970s when the idea of a European monetary union was first developed. During the first decade following the introduction of the euro in 1999, discussions about such transfer systems were all but dead, but it has experienced a revival after the onset of the euro crisis in 2010. This article explains the background and the evolution of the idea.

Keywords

Automatic stabilisers; European Monetary Union; Optimum currency area theory; Unemployment insurance

JEL Classifications

E63; F330; F450

The idea of a fiscal transfer system for the euro area is almost as old as the first proposals for monetary integration of Europe. A number of early reports on the feasibility and preconditions of a common currency in Europe, such as the Marjolin (1975) or MacDougall (1977) reports, include references to and recommendations for fiscal transfer mechanisms between participating countries. Standard textbooks on monetary integration, such as de Grauwe (2014), have also discussed the basic logic of fiscal transfers in currency unions from their first editions onwards.

The Traditional Economic Argument

The basic argument is that with joining a monetary union, a member state gives up some of its key macroeconomic policy instruments, i.e. control over short-term interest rates and the ability to let the domestic currency appreciate or depreciate against other currencies. Compared to flexible exchange rates, this leads to a diminished ability to shelter the domestic economy against external shocks.

As a consequence, the need for alternative adjustment mechanisms arises. One option for such an alternative adjustment mechanism is to increase labour mobility, as explained in Robert Mundell's (1961) seminal work on optimum currency areas. Workers from a country hit by a negative asymmetric demand shock would then migrate to countries which have not been hit by such a shock, rendering devaluations unnecessary.

Another alternative is to deal with asymmetric shocks by using fiscal policies. If a country is hit by a negative asymmetric shock, its government can increase public spending or cut taxes in order to boost domestic demand. However, sometimes a country's government might not have the necessary funds for such a stabilisation policy, especially if shocks are large.

Proposals for fiscal transfer mechanisms between countries (often also dubbed 'insurance against asymmetric shocks') are linked to this argument: if a transfer mechanism is designed in a way that funds are moved from countries being hit by a positive asymmetric demand shock to countries which experience negative shocks, both groups of economies can be stabilised. In the country that experiences the negative shocks, transfers could be used to prop up domestic demand; in the country hit by the positive demand shock, funds available for both private and domestic expenditure would be reduced, thus preventing an overheating of the economy.

Early Proposals

The first detailed mentioning of fiscal transfer mechanisms in the debate on a European monetary union was made in the Marjolin report in 1975. In the annex to this report, the authors discuss the possible institutional details of a 'Community Unemployment Benefit Scheme', proposing to introduce a European fund which pays a certain lump sum to the unemployed, financed by a community contribution on wages. Building on this, for a later point of time, the authors envision the transformation of this insurance with lump sum transfers into an unemployment insurance system more in line with existing national schemes, with benefits based on past earnings.

MacDougall (1977) expands the ideas of the Marjolin report, discussing both a transfer system from a common fund to national unemployment systems dependent on the current cyclical situation as well as a scheme under which only member states in exceptionally dire economic conditions would receive transfers from the rest of the union. Overall, the MacDougall report concludes that for a meaningful macroeconomic stabilisation function, the European budget should be increased to 5–7% of GDP. While this would still have been far below the volume of the federal budget in the USA (which is often taken as a point of reference for the fiscal capacity of the European Monetary Union), it would have been way above what has been reached in Europe even by 2015 (when the EU budget amounted to slightly more than 1% of GDP).

The contributions to this debate further proliferated, with the project of a European monetary union becoming more concrete in the early 1990s. Two main arguments were made during this time.

On the one hand, Asdrubali et al. (1996) and von Hagen (1992) argued that transfer systems in US unemployment did less to bolster state-specific shocks than previously thought (Asdrubali et al. came to the conclusion that US unemployment insurance only stabilised less than 2% of these shocks). On the other hand, the estimates of funds needed for meaningful stabilisation were greatly reduced relative to the volumes referred to in the MacDougall report. Specifically, Italianer and Vanheukelen (1993) demonstrated that a macroeconomic stabilisation comparable to that in the USA could actually be achieved with as little as 0.2% of GDP if intra-regional transfers were linked to the changes of national unemployment rates relative to changes in EU-wide unemployment rates, and hence would be well targeted on the goal of macroeconomic stabilisation.

The Maastricht framework Without Fiscal Transfers

Despite this in-depth debate on fiscal transfer systems, the final framework agreed in the Maastricht Treaty in 1992 did not include any mentioning of such mechanisms. Instead, the Maastricht framework was explicitly built on the idea that each country would be responsible to deal with its own (idiosyncratic) macroeconomic problems. Monetary policy was unified and transferred to the European Central Bank and was sheltered from policy interference both by defining 'price stability' as the ECB's primary goal and by prohibiting monetary financing of government deficits. Fiscal policy was left at the national level. A 'no bail-out clause' actually even prohibited one country or the European Union as a whole from taking over another country's liabilities. In order to prevent spill-overs from unsustainable fiscal policy at the national level to the common monetary policy, the Stability and Growth Pact limited national government deficits to 3% of GDP, except in cases of deep recession.

This reluctance to include automatic stabilisers in the Maastricht framework was likely due to two reasons: one political and one based on changing perceptions of the academic community.

Politically, in the early 1990s there was little appetite to introduce a transnational transfer system with large financial volumes. In particular, the West German electorate had just been burdened with the transfers for rebuilding the East German economy and would not have accepted new transfers to economically weaker euro zone countries.

From an economic (academic) point of view (at that time), new theoretical economic models had led a number of economists to rethink their belief in the need for fiscal transfer systems in order to stabilise national business cycles in a monetary union. First, the advent of New Classical macroeconomic models with rational expectation elements (especially in their real business cycle variety) had led to the conclusion that fiscal policy was all but ineffective for stabilisation of the business cycle. According to these models, macroeconomic fluctuations were created either by exogenous technology shocks or by the reaction of rationally thinking agents to disturbances created by monetary or fiscal policies. As in these models, agents were always in a situation that was optimal given the underlying economic fundamentals; they would react to any anticipated change in fiscal policy and try to counteract its effect, greatly reducing its effectiveness.

In contrast, economists more closely aligned with Keynesian thinking believed that stabilisation through fiscal policy could still be conducted effectively, but transfers would not be a precondition for such policies. Instead, it was assumed that countries with open financial markets could always borrow the resources needed in international markets.

Finally, a number of economists believed that a growing integration of product and factor markets would reduce the need for automatic fiscal stabilisers (von Hagen 1992; see also Bertola and Boeri 2002; Blanchard and Giavazzi 2003). According to these arguments, first, the integration of goods markets would make asymmetric shocks less likely, as shocks would quickly be transmitted to the whole euro area through trade linkages, making them symmetric. Second, with the growing integration of labour markets, increased factor mobility would work as a shock absorber. Third, as Mélitz and Zumer (1999)

pointed out, EMU itself would lead to a closer integration of financial markets and hence provide risk sharing through credit and insurance markets.

Recent Poposals

After the introduction of the euro as a virtual currency in 1999 and in the form of notes and coins in 2002, the discussion of European transfer systems was all but dead. It only re-emerged after the onset of the euro crisis in 2010. As the specifics of the Maastricht Treaty were a result of the specific economic experience of the early 1990s and of the specific economic thinking of the time, the revival of ideas for fiscal stabilisers also can best be explained against the background of change in (mainstream) academic thinking and empirical experiences with the existing institutional setup.

First, theoretically, fiscal policy made a comeback as a stabilisation tool during the 2000s. Extending standard macroeconomic models with frictions such as wage and price stickiness (Galí et al. 2007) or with the assumption of underdeveloped financial markets (Aghion and Howitt 2006) led to the conclusion that the costs of macroeconomic fluctuations were much bigger than previously thought and that fiscal policy was indeed able to reduce volatility and hence increase welfare substantially.

Second, more recent econometric studies have hinted that the US unemployment insurance might have contributed more to macroeconomic stabilisation than previously thought. For example, Chimerine et al. (1999) put the stabilisation effect of US insurance in selected recessions to 15–20% of the initial drop of GDP. Vroman (2010) comes up with a stabilisation effect of almost 30% in the Great Recession of 2008/09. According to Dullien (2014), one reason for the huge differences in these works and the earlier works by Asdrubali et al. or von Hagen et al. are that the earlier research often did not take into account all payment flows in US unemployment insurance and neglected the extended benefits and emergency benefits which are enacted in recessions. Another reason is a difference in the measured concept of stabilisation. While the Asdrubali et al. contribution looks at fluctuations over the whole business cycle, literature in line with Vroman's approach focuses on recessions and compares the actual GDP path to a counterfactual simulation without unemployment benefits. As unemployment only increases strongly in a recession and recessions are rather rare events, stabilisation in a recession is larger than that measured on average over the cycle.

Third, the run-up to and the eruption of the euro crisis have exposed a number of weaknesses of the monetary union's macroeconomic framework. Already, in the years prior to the crisis, large economic divergences within the euro area had been observed (Dullien and Fritsche 2009), which were subsequently linked to the absence of effective macroeconomic stabilisation at the national level. It was argued that housing price bubbles in Ireland and Spain might not have become so large had a transnational transfer system drained purchasing power earlier.

In addition, the euro crisis has cast some countries into extremely deep recessions. These recessions demonstrated that even countries which had a sound budgetary position prior to the crisis might experience problems financing their budget deficits in a downturn. Spain and Ireland were the most obvious examples of this possibility: both countries had actually run budget surpluses prior to the crisis and both countries fought against the fear of insolvency and for continued capital market access during the crisis. This experience cast doubt on the premise that countries could just stabilise the business cycle on their own during a downturn by borrowing in financial markets and spending the funds on tax cuts and public expenditure.

As a consequence, when institutional reform of the euro area was discussed in the wake of the crisis, the topic of a fiscal transfer mechanism re-emerged. Both the European Commission's (2012) 'Blueprint for a Genuine European Monetary Union' and the Four Presidents' Report (drafted by the President of the European Council, the President of the European Commission, the President of the Eurogroup and the President of the European Central Bank) mentioned explicitly the long-term goal of transnational fiscal transfers.

According to the latter, European policymakers should build 'a well-defined and limited fiscal capacity to improve the absorption of country-specific economic shocks, through an insurance system set up at the central level' (van Rompuy 2012, p. 5). While the four presidents remained relatively vague on the details of the desired fiscal capacity, they explicitly discussed, as possible options, both a cross-country insurance system with transfers between national budgets and a European unemployment insurance with individual claims to benefits. In a related paper, the European Commission (2013) later outlined a vision for the 'Strengthening of the Social Dimension of Economic and Monetary Union' in which a European unemployment insurance was discussed in detail. However, the discussion in the so-called 'Five Presidents' Report' of 2015 (a follow-up report to the Four Presidents' Report published under a new Commission and with the inclusion also of the president of the European Parliament) on this topic was subsequently toned down (Juncker et al. 2015). While still asking for a 'euro area stabilization function' through fiscal transfers, the five authors avoided providing any clear direction in the debate and only defined some required characteristics of such a system – among others that it must not create permanent transfers and not distort incentives for national policy makers for sound fiscal policymaking or structural reforms.

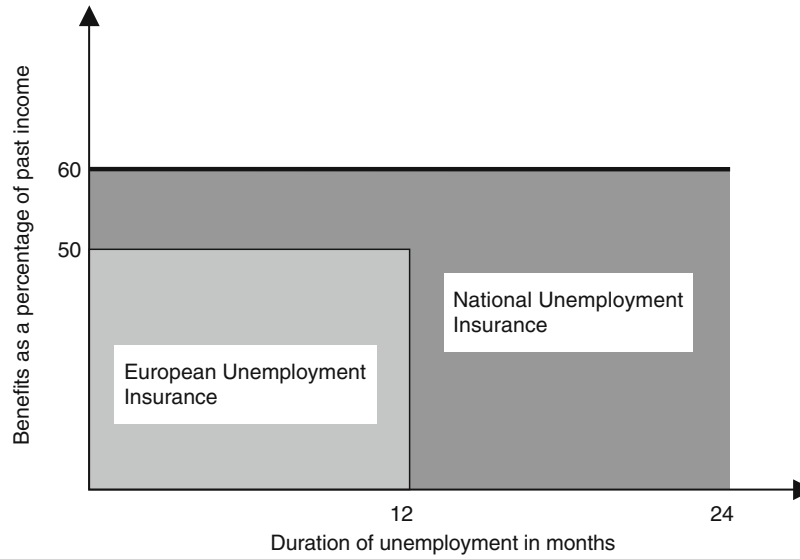
From the academic side, this discussion was supported by a number of more or less elaborate proposals for the design of automatic stabilisers for the euro area. These new contributions can be grouped into three categories: (1) proposals for a genuine unemployment insurance system that directly provides some kind of transfers to the unemployed; (2) proposals for reinsurance schemes which transfer money into national unemployment insurances in case of financial strain; and (3) proposals for transfer systems between national budgets.

Proposals for Genuine Unemployment Insurance Systems

The most prominently discussed proposal for genuine unemployment insurance is the one originally presented in Dullien (2007, 2008), although Deinzer (2004) had presented a similar proposal slightly earlier. According to Dullien's proposal, all employees in the euro area would pay a small payroll tax on their wages up to a certain threshold into a European unemployment insurance. After a certain number of months of insured employment (e.g. 20 out of the past 24 months), an individual would have the right to receive unemployment benefits in the event of job loss for a limited period of time. The benefits would be set as a share of past earnings.

The benefit duration and replacement rates from the European system would be defined roughly in line with the lowest common denominator for existing systems in euro area countries. In addition, each country would be allowed to top up the European system with national means, with regard to its duration, replacement rate and duration of unemployment benefits, to reach levels similar to those before the introduction of a European unemployment insurance. The aim of this setup would be to leave individual incentives to look for a new job unaltered by the introduction of this European unemployment insurance and allow for different levels of social protection between the euro area member states, depending on national preferences. Figure 1 illustrates this principle: prior to the introduction of a European unemployment insurance, all of the replacement payments are paid by the national unemployment insurance. From the individual unemployed's perspective, the solid black line represents the replacement payments. After the introduction of a European unemployment insurance, this transnational system pays 50% of past earnings for one year, while the national system would top up this payment for the first 12 months and cover the whole payment afterwards. In the figure the light grey area is paid by the European system and the dark grey area by the national system. From the individual unemployed's perspective, the solid black line would represent the replacement payments.

It is important to note that in contrast to national unemployment schemes, which were mostly introduced with social objectives in mind (see e.g. for the US case Blaustein 1993), the motivation for a European unemployment



European Unemployment Insurance, Fig. 1 Principle of a top-up unemployment insurance

insurance is mostly macroeconomic. Almost all euro area countries already have unemployment insurance systems in place which are actually more generous than the proposed European scheme, and hence no overall improvement in replacement rates for the unemployment would be expected from the introduction of such a scheme. Only the flow of funds would change: after the introduction of the European unemployment insurance a significant share of replacement payments would run through the European system instead of through national unemployment systems and national budgets. As national public budgets in the euro area are constrained through various fiscal rules (and sometimes by markets cutting off countries from the possibility to borrow at acceptable rate), this would give national governments more fiscal breathing space in recessions and hence hopefully lead to more countercyclical fiscal policies.

Proposals for Reinsurance Systems

In contrast to these genuine unemployment insurance systems, reinsurance systems such as the one proposed by the Brussels-based Centre for Economic Policy Studies (Beblavý and Maselli 2014) only transfer money to countries which experience a deep recession. Moreover, instead of insuring individuals (as in the genuine unemployment systems), here the insured entity would be the existing national unemployment insurances.

Specifically, the CEPS economists propose that each country pays 0.1% of its GDP annually into a European buffer fund until the fund's reserves reach 0.5% of euro area GDP. If, now, unemployment in one member state increases by more than two percentage points above its equilibrium unemployment rate, the country's unemployment insurance receives resources from the fund. It is further envisioned that countries that have drawn an excessive amount from the system would see their contribution rate increased to 0.2% of GDP.

The logic of this system is based on the reasoning of insurance economics: small damages (low to medium rates of unemployment) are assumed to be manageable by member states alone, while for large damages reinsurance payments are considered necessary. The contribution here is seen as an 'insurance premium' which is adjusted according to the risks insured, with high-unemployment countries paying a higher premium.

Proposals for Transfer Systems Between National Budgets

The third group of proposals, recommending rule-based direct transfers between national governments' budgets, is best represented by Enderlein

et al.'s (2013) idea of a 'cyclical shock insurance'. According to this proposal, euro area countries would pay into a transfer system and receive funds from this system depending on their position in the economic cycle relative to other euro area members. To measure the position in the business cycle, Enderlein et al. propose using the output gap (defined as the deviation between potential and current output), as this variable is already used in other euro area policy contexts (e.g. in the evaluation of excessive budget deficits).

According to this proposal, if a country's output gap is more negative (or less positive) than the euro area's average (or current output further below or less above potential output), it receives funds from the common pool. If a country's output gap is less negative (or more positive) than the currency union's average (or current output further above or less below potential), it would pay into the common pool. According to the Enderlein et al. proposal, the system's finances would be balanced each year, even though one could envision a system with direct transfers between national budgets without this specific feature.

Stabilisation Impact of Different Schemes

While the systems look rather similar at first sight, they have very different stabilisation properties. Firstly, some of the systems can help in stabilising the national and the overall business cycle *over time*. This is the case if the system is allowed to receive more funds in good times than it pays out, and pays out more in bad times than it receives – in other words it is allowed to accumulate reserves and go into deficit. Such a property would lead to a transfer of aggregate demand from boom times to recessions and the overall business cycle would be dampened.

In contrast, some of the systems might *only stabilise across space*. This is the case when the system's finances have to be balanced each year. Here, funds are transferred only between countries that do relatively better than others towards those that do relatively worse. While this kind of system might help to bring the euro area's business cycles closer together (and hence make a single monetary policy more appropriate for each single member state), it might actually *destabilise* the national business cycle. For example, in the deep recession in the global financial and economic crisis of 2008/09, all euro area countries saw output falling. Yet GDP in some countries contracted by more than in others. Systems which only stabilise between countries would have taken funds from countries with only a small decrease in GDP (but still in recession) and transferred them to the countries more severely hit, further decreasing GDP (and hence deepening the recession) in the former, but increasing GDP (and alleviating the recession) in the latter. This feature of these systems is an unavoidable consequence if they are not allowed to go into deficit or accumulate reserves. For the supporters of these schemes, this is usually seen as a secondary problem as they (implicitly or explicitly) rely on the European Central Bank to stabilise the overall business cycle and hence see the aim of the fiscal stabilisation scheme only as increasing convergence of the national business cycles.

A third question is in how far the systems work symmetrically: that is, whether they also stabilise the business cycle in a boom. If a system relies on fixed contributions, but pay out only in bad times, it will not significantly dampen an economic boom. In contrast, if contributions and/or payouts vary proportionally over the whole business cycle, booms are also dampened. The question of symmetric stabilisation is important if one believes that one of the reasons for recent deep recessions in some euro area member states was a prior overheating of national economies, as is often argued in relationship with the construction boom in Spain prior to the crisis of 2008/09.

Table 1 summarises the three dimensions of stabilisation for three representative proposals mentioned above. Looking at stabilisation over time and over space, genuine unemployment insurance stabilises in both dimensions: firstly, it stabilises both the overall euro area business cycle and national business cycle, as payouts are higher in a recession than in a boom and contributions are higher in a boom than in a recession. This holds

European Unemployment Insurance, Table 1 Stabilisation properties of different proposals for fiscal transfer systems

	Stabilisation of the overall (also national) business cycle	Stabilisation between countries (narrows the deviation of output gaps among euro area member states)	Symmetric stabilisation – dampens (relative) booms as well as (relative) busts
Genuine unemployment insurance	++	+	+
Reinsurance for national unemployment insurance systems	++	+	0
Cyclical shock insurance	–	++	++

+ some stabilisation; + + strong stabilisation; 0 no stabilisation; – potential destabilisation (Source: Author's elaboration)

both at a European level for the overall business cycle and at the national level for the national business cycle. Secondly, it also stabilises between countries, as countries that do relatively better pay in more (or receive less) than countries doing less well. Thirdly, the genuine unemployment insurance works symmetrically in a boom and a bust. In a boom, growing wages and employment rates lead to higher contributions and hence less disposable income in the country in question. In a bust, a country receives more unemployment benefits.

For the reinsurance model, similar arguments apply when it comes to the question of whether the overall business cycle is stabilised and whether cross-country differences in the business cycle are reduced: as the system provides assistance in a downturn and does so to all countries that have unemployment rising above the threshold, it stabilises a country's (and the euro area's) overall business cycle. As countries that are harder hit receive larger payments, it also moves the cyclical position of countries closer together. Yet it does so only when it is activated (in recessions). Reinsurance also differs in the question of symmetry of stabilisation: as payouts are only made in a crisis and contributions are fixed over the cycle, it does not stabilise in a boom, but only in a recession.

In contrast, the cyclical shock insurance does not stabilise a country's overall business cycle and it might even deepen recessions or magnify booms in single member states when the whole euro area is in recession or in a boom, as it only redistributes between countries but is not designed to accumulate or run down reserves. This potential destabilisation in a recession is not only an economic issue, but also a political economy problem: it is very likely that a system which forces a country already in recession (but in a less deep recession than the euro area as a whole) to further cut expenditure and transfer the funds to the European system (thus deepening the national recession) would quickly become the target of public criticism in the country concerned.

On the positive side, the cyclical shock insurance has the strongest effect for cross-country stabilisation and also provides the strongest symmetric stabilisation among the proposals: by its construction, it takes as many funds from booming economies as it pays out to weak economies.

Other Advantages and Disadvantages of the Proposed Schemes

Beyond the stabilisation impact, there are a number of other advantages and disadvantages of the current proposals which have been discussed.

For genuine unemployment insurance, the potentially complicated implementation, given the interaction with different, already complex existing national unemployment insurance systems, is often mentioned. In addition,

administering the database of contribution history of all of the insured in participating countries would require the creation of a significant central infrastructure, which would be more expensive than the rather sleek administrations necessary for just calculating transfers between national unemployment insurances or national budgets based on macroeconomic indicators.

Moreover, given limitations of data availability on the individual employment and earning histories of the unemployed, it is not currently possible to simulate the exact payment flows that such a scheme would create. Thus there is a risk that the system would create permanent transfers. An attempt to simulate a basic European unemployment insurance with the EUROMOD model came to the conclusion that, for the period from 2000 to 2013, some countries would have become permanent net payers or net recipients (Dolls et al. 2014). As Bargain et al. (2013) show for the case of the introduction of a common tax and benefit system for Europe, these transfers might have surprising directions (in their simulation, Germany becomes a net recipients of transfers). While, in principle, such a risk could be mitigated by allowing contribution rates to vary between countries if single member states draw too much from the system, such a variation in contribution rates also carries the risk of lowering the stabilisation impact of the scheme.

A related issue is the question of how far the introduction of European unemployment insurance creates a moral hazard for national governments. One issue here is in how far such a system might alter the incentives for national governments to conduct structural reforms. Occasionally, it is argued that the introduction of such a system would reduce national governments' willingness to liberalise their labour markets, as some of the costs of unemployment are moved to the European level. Yet this argument is not entirely convincing. Firstly, by its construction, the genuine unemployment insurance only covers short-term unemployment (as the unemployed only receive benefits after a substantial period of insured employment and only for a very limited time). Labour market reforms usually aim at reducing structural unemployment, which by its nature is long-term employment. Secondly, many structural reforms discussed for European labour markets (e.g. the loosening of dismissal rules) would initially lead to more short-term unemployment, not less. Thus, if part of these costs were borne at the European level, the incentives of national governments for such structural reforms might actually increase.

A further moral hazard issue is the question of how far national governments or administrations could minimise national net payments into the system. As has been observed in other federal systems, lower levels of the administration might reduce their efforts to put the unemployed to work if higher levels are responsible for paying the related unemployment benefits. Thus, it would be possible that national unemployment administrations might focus more strongly on putting those unemployed who receive benefits exclusively from the national system (e.g. because the minimum requirements for receiving European benefits have not been met) into new jobs and neglect those who are mostly financed by the European insurance. It is unclear, however, how big a problem this really would be as, empirically, most short-term unemployed who find a new job do so on their own and not through the unemployment administrations' efforts.

On the positive side, a clear advantage of genuine unemployment insurance would be that the concept of European unemployment insurance could be easily communicated to the general public, as most countries already have a similar national system of unemployment insurance in place. It has also been claimed that European unemployment insurance could provide a 'human face' for the European integration project (Fattibene 2015).

When it comes to the reinsurance system, one potential problem is the binary character of the potential triggers. As payments are only made if unemployment rises above a certain threshold, but then payments might be large, there might be an incentive for national governments to push unemployment beyond this threshold if it is already close. However, this criticism could be mitigated if payouts were phased in above a certain threshold instead of being allowed to kick in fully after the threshold is reached.

Another issue is the question of how far the unemployment rate is an appropriate indicator for determining transfers. Only part of the unemployment observed in Europe is of a cyclical nature, while in many countries the larger part of unemployment is considered to be structural. This is why the CEPS authors propose linking payments to deviations of the national unemployment rate from the structural unemployment rate (technically, the non-accelerating wage rate of unemployment: NAWRU). Yet, as discussed in Blanchard and Katz (1997) and other contributions to the winter 1997 issue of the *Journal of Economic Perspectives*, all real-time empirical estimations of the natural rate of unemployment are subject to a large degree of uncertainty, so it is possible that under reinsurance proposals payments are linked to questionable underlying data. However, one should keep in mind that, due to the specific construction of the payments to a trigger value, this problem is mitigated: even if the estimate of the underlying natural unemployment rate is off by one percentage point, countries meeting the trigger value of their actual unemployment rate having risen by two percentage points above the estimated national rate are clearly in a recession and would benefit from the stabilisation.

A third question – which applies equally to cyclical shock insurance – is how to make sure that the funds disbursed to national unemployment insurance systems or national budgets are also spent in a timely manner, which is a precondition for effective business cycle stabilisation. The problem is that, given the trigger issue, national governments faced with a moderate recession would initially have to plan their budgets without revenues from the central system if they do not want to risk excessive deficits. If, during the year, unemployment worsens and payouts are made from the reinsurance, it might be difficult to spend the extra funds quickly in a meaningful way.

In addition to the problem that the cyclical shock insurance might exacerbate national recessions under certain conditions (as discussed above), it is also often criticised for its reliance on the output gap for the calculation of transfers. While the output gap is a straightforward theoretical concept (defined as the difference between potential and current output), its empirical measurement is fraught with difficulties. Usually, econometric filter techniques such as the Hodrick–Prescott filter are used to estimate potential output. The problem is that the end point (i.e. the latest estimate in a time series) is susceptible to large changes if future output turns out differently than assumed. As a consequence, estimates of potential output are often very strongly revised *ex post* and, as a consequence, estimates for the output gap are revised even more strongly. Especially in times of crisis, these revisions tend not only to be large, but also to change the sign of the estimated output gap (Kempkes 2012). These revisions might result in payments which actually amplify the business cycle, rather than stabilising it. The European Commission's estimates for the Spanish output gap during the late 2000s are a case in point. In spring 2007, the Commission put the Spanish output gap for 2006 at minus 1.1% of GDP, meaning that current output was 1.1% below potential. Without large revisions of the underlying GDP data, but just because of new knowledge about GDP in the years 2007 and after, the Commission's estimate for the output gap 2006 from Autumn 2012 stood at *plus* 1.8%, meaning now that current output in that year was seen as being 1.8% in excess of potential output. Had one made transfers according to the real-time estimates, Spain would have received extra stimulus at a time when (at least with hindsight) the economy was already overheating.

Similar to the question of a potential destabilisation of the national business cycle, this is not purely an economic problem, but also a political economy one: it is difficult to imagine that a system under which a handful of econometricians in Brussels calculate large transfers between national budgets on a questionable methodological basis will be accepted by the taxpayers (and voters) of potential net payers.

For both the reinsurance system and the cyclical shock insurance, the low number of staff needed to administer the scheme is often mentioned as an advantage: only aggregate national data need to be analysed, so there is no need for a large central administration.

Political Considerations

In addition to these economic considerations, there are a number of political obstacles to overcome before (if ever) European unemployment insurance or one of the other fiscal stabilisation systems can be implemented. While it is disputed which of the proposals would require a treaty change (and hence unanimity), it is clear that the introduction of any fiscal stabilisation scheme would only be possible if there were a broad consensus among euro member states for the proposed scheme. At the time of writing, such a consensus is not in sight.

Predictably, all of these proposals are challenged by those who prefer limiting the EU's reach into national sovereignty. In addition, the single proposals each face political challenges of their own.

The genuine unemployment insurance faces the largest political resistance for a number of reasons. Firstly, the common European fund would be allowed to go into deficit in a deep recession (especially if a recession were to hit right after the fund's introduction); it might be seen by some as the introduction of 'euro bonds through the back door' and hence is faced with opposition from opponents of joint liabilities. Moreover, as this proposal carries the largest risk of creating permanent transfers (as exact payments are difficult to simulate *ex ante* with currently available data), it is opposed by many of the countries which saw their unemployment increase only a little in the crisis of 2008/09 and the subsequent euro crisis (and which hence would have been net payers during this time), such as Germany or the Netherlands. Thirdly, some countries have long opposed increasing EU competencies for social policies and hence fear that the introduction of genuine unemployment insurance at the European level might be used by EU institutions to limit national discretion for social protection. Finally, as the unemployment insurance systems in many euro area countries are managed under the strong influence of social partners, the national unions and national employers' federations fear the loss of influence if a significant part of the funds were in future funnelled through a European system.

The cyclical shock insurance faces the problem of shaky methodological foundations of the estimates of the output gap. While some authors have put some hope into improving the estimation methods for the output gap (Carnot et al. 2015), the underlying methodological problem – notably that it is difficult to estimate a trend in real time when the future is uncertain – will not go away. While there are not many interest groups which feel directly threatened by cyclical shock insurance, the technical issues will make it difficult to communicate the idea to the broader public and hence gain backing by national voters for the proposal.

The reinsurance system seems to be confronted with the least political resistance: it remains relatively easy to explain and yet it does not threaten directly specific interest groups. Yet so far very few politicians on the national level have put their weight behind the proposal.

In conclusion, one can see that, even though the number of supporters for some kind of fiscal transfer system for the euro area in order to stabilise the business cycle of the currency union has been growing, and there is a certain consensus about the desirability of additional macroeconomic stabilisation, none of the stabilisation mechanisms proposed is perfect and ready to be implemented. Even if political consensus about the introduction of such a scheme could be reached among member states (which seems to be difficult enough at the moment), a lot of work would have to be put into the technical details of these proposals before they truly could become operational.

See Also

- ▶ European Monetary Union
- ▶ Euro Zone Crisis 2010
- ▶ Genuine Economic and Monetary Union

Bibliography

Aghion, P., and P. Howitt. 2006. Joseph Schumpeter Lecture: Appropriate growth policy: A unifying framework. *Journal of the European Economic Association* 4: 269–314.

Asdrubali, P., B.E. Sorensen, and O. Yosha. 1996. Channels of interstate risk sharing: United States 1963–1990. *Quarterly Journal of Economics* 111(4): 1081–1110.

Bargain, O., M. Dolls, C. Fuest, D. Neumann, A. Peichl, N. Pestel, and S. Siegloch. 2013. Fiscal union in Europe? Redistributive and stabilizing effects of a European tax-benefit system and fiscal equalization mechanism. *Economic Policy* 28: 375–422.

Beblavý, M., and I. Maselli. 2014. An unemployment insurance scheme for the euro area: A simulation exercise of two options. CEPS Special Report No. 98. Brussels.

Bertola, G., and T. Boeri. 2002. EMU labor market two years on: Microeconomic tensions and institutional evaluation. In *EMU and economic policy in Europe: The challenge of the early years,* ed. M. Buti and A. Sapir. Cheltenham: Edward Elgar.

Blanchard, O., and F. Giavazzi. 2003. Macroeconomic effects of regulation and deregulation in goods and labor markets. *Quarterly Journal of Economics* 118(3): 879–907.

Blanchard, O., and L.F. Katz. 1997. What we know and do not know about the natural rate of unemployment. *Journal of Economic Perspectives* 11(1): 51–72.

Blaustein, S.J. 1993. *Unemployment insurance in the United States: The first half century.* Kalamazoo: W. E. Upjohn Institute for Employment Research.

Carnot, N., P. Evans, S. Fatica, and G. Mourre. 2015. Income insurance: A theoretical exercise with empirical application for the euro area. European Commission Economic Papers 546. Brussels.

Chimerine, L., T.S. Black, and L. Coffey. 1999. Unemployment insurance as an automatic stabilizer: Evidence of effectiveness over three decades. Unemployment Insurance Occasional Paper 99-8.

de Grauwe, P. 2014. *Economics of monetary union*, 10th ed. Oxford: Oxford University Press.

Deinzer, R. 2004. *Konvergenz-und Stabilisierungswirkungen einer europäischen Arbeitslosenversicherung*. Berlin: Duncker & Humblo.

Dolls, M., C. Fuest, D. Neumann, and A. Peichl. 2014. An unemployment insurance scheme for the euro area? A comparison of different alternatives using micro data. ZEW Discussion Paper No. 14-095. Mannheim.

Dullien, S. 2007. Improving economic stability in Europe: What the euro area can learn from the United States' unemployment insurance. Working Paper Stiftung Wissenschaft und Politik FG 1, 11. Also available online at www.swpberlin.org/fileadmin/contents/products/arbeitspapiere/Paper_US_KS_neu_formatiert.pdf. Accessed 7 July 2015.

Dullien, S. 2008. *Eine Arbeitslosenversicherung für die Eurozone: Ein Vorschlag zur Stabilisierung divergierender Wirtschaftsentwicklungen in der Europäischen Währungsunion*. Berlin: SWP-Studien S01.

Dullien, S. 2014. *A European unemployment benefit scheme: How to provide for more stability in the Euro Zone*. Gütersloh: Bertelsmann Stiftung.

Dullien, S., and U. Fritsche. 2009. How bad is divergence in the euro zone? Lessons from the United States and Germany. *Journal of Post Keynesian Economics* 31(3): 431–457.

Enderlein, H., L. Guttenberg, and J. Spiess. 2013. Making one size fit all. Designing a cyclical adjustment insurance fund for the euro zone. Notre Europe policy Paper 61, Berlin.

European Commission. 2012. *A blueprint for a deep and genuine economic and monetary union*. Brussels: European Commission.

European Commission. 2013. *Strengthening of the social dimension of economic and monetary union*. Brussels.

Fattibene, D. 2015. Creating a union with a "human face": A European unemployment insurance. Instituto Affari Internazionale Working Paper 15/13. Rome.

Galí, J., M. Gertler, and D.J. Lopez-Salido. 2007. Markups, gaps, and the welfare costs of business fluctuations. *Review of Economics and Statistics* 89(1): 44–59.

Italianer, A., and M. Vanheukelen. 1993. Proposals for Community stabilization mechanisms: Some historical applications. European Economy, Reports and Studies No 5: 493–510. Brussels.

Juncker, J.-C., D. Tusk, J. Dijsselbloem, M. Draghi, and M. Schulz. 2015. *Completing Europe's economic and monetary union*. Brussels: European Commission.

Kempkes, G. 2012. *Cyclical adjustment in fiscal rules: Some evidence on real-time bias for EU-15 countries*. Deutsche Bundesbank Discussion Paper No. 15/2012. Frankfurt am Main: Deutsche Bundesbank.

MacDougall, D. 1977. *Report on the study group on the role of public finance in European integration, chaired by Sir Donald MacDougall*. Economic and Financial Series No. A13. Brussels: Commission of the European Communities.

Marjolin, R. 1975. *Report of the study group "Economic and monetary union 1980"*. Brussels: Commission of the European Communities, Directorate-General of Economic and Financial Affairs.

Mélitz, J., and F. Zumer. 1999. Interregional and international risk-sharing and lessons for EMU. *Carnegie-Rochester Conference Series on Public Policy* 51(1): 149–188.

Mundell, R.A. 1961. A theory of optimal currency areas. *American Economic Review* 51(4): 457–480.

van Rompuy, H. 2012. *Toward a genuine economic and monetary union. Report by the President of the European Council*. Brussels: European Council.

von Hagen, J. 1992. Fiscal arrangements in a monetary union: Evidence from the U.S. In *Fiscal policy, taxation and the financial system in an increasingly integrated Europe,* ed. D.E. Fair and C. de Boissieu. Dordrecht: Kluwer.

Vroman, W. 2010. *The role of unemployment insurance as an automatic stabilizer during a recession*. Washington, DC: Urban Institute.

European Union (EU) Research and Experimental Development (R&D) Policy

Henri Delanghe

Abstract

The European Union (EU)'s research and experimental development (R&D) policy pursues a range of objectives stated explicitly in the Treaty. Over time, EU R&D policy has gained in importance compared with other EU policies (as a result of the EU Lisbon, revised Lisbon and Europe 2020 strategies) and compared to Member State policies (as a result of a better recognition of the added value of action at EU level). EU R&D policy has so far consisted mainly of supply side-oriented direct financial support in the form of ever larger and more complex multi-annual 'Framework Programmes' supporting cross-border research programme coordination, frontier and cross-border collaborative research projects, international researcher mobility, research infrastructure access etc. Large-scale 'additional' impacts have been achieved by these Framework Programmes. In recent years, indirect support in the form of policy advocacy has gained in importance.

Keywords

Economic policy; EU framework programme; European research council; Horizon 2020; Lisbon strategy; Lisbon treaty

JEL Classifications

O3; O31; O320; O380

Disclaimer All views expressed herein are entirely of the author, do not reflect the position of the European Institutions or bodies and do not, in any way, engage any of them.

The General Framework of R&D and Technological Innovation Policy

The basic framework for R&D policy has already been described elsewhere (see 'R&D and technological innovation policy'). R&D, when appropriately valorised, leads to technological innovation in the form of new products and processes, which contribute to growth, competitiveness and job creation, and which produce other societal benefits. Because of market failures, the private sector, left to its own devices, invests in R&D in sectors not always fully aligned with, and at levels below, the socially desirable, and is unable to fully valorise its research output, which justifies public intervention. The latter needs to be thought through carefully, based on *ex ante* impact assessment informed by credible *ex post* evaluation.

The Stated Objectives of EU R&D Policy

The stated objectives of the EU's R&D policy fully recognise the potential societal impacts of R&D. As revised most recently under the Lisbon Treaty, they read as follows: 'The Union shall have the objective of strengthening its scientific and technological bases by achieving a European research area in which researchers, scientific knowledge and technology circulate freely, and encouraging it to become more competitive, including in its industry, while promoting all the research activities deemed necessary by virtue of other Chapters of the Treaties' (for the consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union and the Charter of Fundamental Rights of the European Union, please refer to the website: http://register.consilium.europa.eu/pdf/en/08/st06/st06655-re07.en08.pdf).

The Increasing Importance of R&D Policy Compared to other EU Policies: The Lisbon, Revised Lisbon and Europe 2020 Strategies

Over time, the EU's R&D policy, i.e. the policy put in place to achieve the aforementioned Treaty

objectives, has gained in importance. This trend has accelerated over the past 10–15 years, mainly as a result of two factors. The first is the move of R&D policy to the heart of the EU policy agenda (just as R&D policy has also gained a central place in national policy agendas over the same period). The Founding Treaties had already provided the Community with a responsibility in the field of research, yet research policy remained mainly a national affair until at least the late 1970s. European research policy was of an *ad hoc* nature, tied to particular sectors (agriculture, coal, nuclear energy, steel etc.) and fragmented. The inclusion of a separate chapter on R&D in the Single European Act (1986) heralded the shift towards a legally solidly grounded integrated European research policy focusing on the competitiveness of European industry and the quality of life of European citizens. Yet it was only as a result of the knowledge-based so-called 'Lisbon' (2000) and 'revised Lisbon' (2005) strategies that R&D policy moved to the heart of EU policy.

The purpose of the 'Lisbon Strategy' was 'to become the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion' (Presidency Conclusions, Lisbon European Council, 23 and 24 March 2000). An important supporting objective in the field of R&D was the so-called 'three per cent' objective. In 2002, the Barcelona European Council agreed that 'overall spending on R&D and innovation in the Union should be increased with the aim of approaching 3 per cent of GDP by 2010. Two thirds of this new investment should come from the private sector' (Presidency Conclusions, Barcelona European Council, 15 and 16 March 2002). Because of the perceived lack of success of the 'Lisbon Strategy', a 'revised Lisbon Strategy' was defined which focused more on growth and jobs and identified the following areas for priority action: (1) investing more in knowledge and innovation; (2) unlocking business potential, especially for SMEs; (3) increasing employment opportunities for priority categories; and (4) climate change and energy policy for Europe (Spring European Council 25–26 March 2005). The successor to the Lisbon Strategies, i.e. the 'Europe 2020' strategy (2010) currently being implemented, continues in this vein with its basic goal of smart, sustainable and inclusive growth.

The Increasing Importance of EU R&D Policy Compared to Member State R&D Policies: The Concept of european Added Value

The second factor driving the increasing importance of EU R&D policy is the more explicit consideration and recognition by the Member States of the added value that can be produced at European level. The concept of European added value relates to one of the key principles underpinning the EU Treaty, the subsidiarity principle, which states that in areas of shared competence, like R&D, the EU shall act only if and in so far as it adds value, i.e. can achieve proposed objectives better than the Member States. In this respect, Article 5 of the consolidated version of the Treaty on European Union states that 'the use of Union competences is governed by the principles of subsidiarity and proportionality' and that 'under the principle of subsidiarity, in areas which do not fall within its exclusive competence, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level' (see the consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union and the Charter of Fundamental Rights of the European Union).

Because of its key role in the Treaty and its increasing political visibility, the added value of action at EU level cannot be automatically assumed and needs to be demonstrated explicitly. This requirement is set out formally in the Treaty Protocol on the Application of the Principles of Subsidiarity and Proportionality, which states that 'the reasons for concluding that a Union objective can be better achieved at Union level shall be substantiated by qualitative and, wherever

possible, quantitative indicators' (see the consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union and the Charter of Fundamental Rights of the European Union). This obligation is also included in the European Commission's guidelines on *ex ante* impact assessment (for an overview of EU IA, please refer to the website: ec.europa.eu/governance/impact). *Ex ante* impact assessment is obligatory for all major new EU initiatives and consists of assessing the applicability of the subsidiarity principle, in addition to assessing the future economic, social and environmental impacts. The former is accomplished by answering the following set of questions:

- Why can the objectives of the proposed action not be achieved sufficiently by Member States (necessity test)?
- As a result of this, can objectives be better achieved by action by the Community (test of EU Value Added)?
- Does the issue being addressed have transnational aspects which cannot be dealt with satisfactorily by action by Member States? (e.g. reduction of CO_2 emissions in the atmosphere)
- Would actions by Member States alone, or the lack of Community action, conflict with the requirements of the Treaty? (e.g. discriminatory treatment of a stakeholder group)
- Would actions by Member States alone, or the lack of Community action, significantly damage the interests of Member States? (e.g. action restricting the free circulation of goods)
- Would action at Community level produce clear benefits compared with action at the level of Member States by reason of its scale?
- Would action at Community level produce clear benefits compared with action at the level of Member States by reason of its effectiveness?

In the field of research, the issue of European added value arose at a relatively early stage. In 1973, the then European Commissioner for Research, Science and Education, Ralf Dahrendorf, asked the following set of targeted questions, all of them still relevant and not fully answered today:

> Are there points at which competition between national science policies – in itself a healthy phenomenon – makes no sense or has not got the desired effect? Are there subjects of scientific inquiry which require, because of their order of magnitude or for other reasons, a co-operative European effort? Are there needs for co-operation across borders due to the specific nature of given problems? Are there efforts of science for which the competitive space is worldwide, with Europe as one competitor rather than European with internal competition? Should co-operation take the form of co-ordination or of a common research effort? Since all this may sound rather abstract, can we afford to have several publicly financed research programmes for the development of telecommunications? Should there be a European computer programme? Can we pool our resources to conduct research into pollution and methods to fight it? Are there experiences in research in occupational medicine which should be shared? It is not difficult to prolong this list of subjects which require public support and have relevance beyond the borders of any single country: methods of enriching uranium, fighting epidemics, urban planning, identifying regional policy needs, producing software for computers, providing European programmes of management training, etc. Thus I should like to be able to take it for granted that there is a place for a European science policy (Dahrendorf 1973).

This explains why, compared with other policy areas, exploration of the issue is relatively advanced in this field. Since 1987 there have been successive multi-annual EU Framework Programmes for Research, Technological Development and Demonstration, the design of which necessarily involves identifying areas and means of EU intervention that would deliver European added value. Since the 1990s, there have been a number of large-scale *ex post* evaluations of these Framework Programmes. While these have acknowledged the evolving nature of the concept of European added value, they have stressed the need for clear *ex ante* definitions, i.e. the robust *ex ante* identification of instances in where the EU can add value. Good examples of such early calls are the so-called 'Five-Year Assessment 1995–1999' (Majó et al. 2000), which concluded that 'there is a need for a better understanding of European Added Value and for a more precise and

operationally useful definition of this concept', and the 'Five-Year Assessment 1999–2003' (European Commission 2005), which concluded that 'a simple and robust definition of European Added Value is needed for the design and implementation of future Framework Programmes'.

In response, Framework Programme evaluators and the European Commission have attempted to compile *ex ante* lists of instances of European added value. The 'Five-Year Assessment 1992–1997' (European Commission 1997) made a start by concluding that 'evidence of European added value is demonstrated by: the existence of important large-scale facilities which no individual Member State would develop and sustain; the promotion of internationally competitive R&D communities in new interdisciplinary areas such as information technology and biotechnology; the creation of strong European industrial platforms based on common technical standards able to compete or cooperate at a global level e.g. mobile telecommunications; the development of pan-European norms and standards for commercial applications'. Stampfer thought European added value was clearer in the case of, for instance, big infrastructures, 'single issue instruments', and European Research Area coordination instruments like ERA-NETs (Stampfer 2008). The most recent attempt to define in an *ex ante* manner instances of European added value in the field of research was made in the *ex ante* impact assessment of Horizon 2020, the new (2014–2020) Framework Programme for Research and Innovation (see "Appendix"; European Commission 2013, Box 4 and Annex 2).

The Evolution of the Framework Programmes

The main policy tool for implementing the EU's R&D policy so far has been direct financial support disbursed through the aforementioned multi-annual Framework Programme. There have been seven Framework Programmes so far (FP1: 1984–1987; FP2: 1987–1991; FP3: 1990–1994; FP4: 1994–1998; FP5: 1998–2002; FP6: 2002–2006; FP7: 2007–2013) and the eighth (2014–2020), called Horizon 2020, has just started (for detailed bibliographical references for the remainder of this section, please refer to Muldur et al. (2006, Chapter 4) and European Commission (2013, Annex 1)).

In accordance with the aforementioned increasing importance of the EU's R&D policy, the FPs have experienced large-scale budget increases over time, from about h5 billion for FP1 to about €70 billion for Horizon 2020. Because of these increases, the EU R&D policy budget has increased in importance not only compared to other EU policy budgets but also to Member State R&D policy budgets, in particular the competitively allocated parts of those. FP5, for instance, which preceded the much larger programmes FP6, FP7 and Horizon 2020, already accounted for a quarter of total funding for publicly financed research projects in the EU.

Most FPs so far have focused on the supply side through subsidies for, for instance, cross-border research programme coordination, frontier and cross-border collaborative research projects, international researcher mobility, research infrastructure access, etc. Horizon 2020 is the first FP to consider the demand side as well, through support for public procurement, for instance. Most of the FPs so far have also focused on research at the expense of innovation. FP6 was the first FP to include innovation support measures. FP7 no longer included such measures, however. Learning the lessons from the past, Horizon 2020 will once more bring research and innovation into a single programme.

In accordance with the expanding range of activities understood to be marked by European added value, the scope of the FP has expanded over time and its structure has become more and more complex: from one FP to another, new activities have been added on, such as public–public partnerships under FP6, and the European Research Council, public–private partnerships and the so-called Risk-sharing Finance Facility under FP7.

Public–public partnerships promote the coordination of public research funding at Member State level. These initiatives were developed based on a double observation: that, on the one hand, apart from the (supranational) FP and a number of inter-governmental initiatives, most

(i.e. well over 80% of) public funding for research in Europe remains uncoordinated and that, on the other hand, coordination can generate benefits in the form of improved programme scope and programme depth, greater scientific excellence because of increased competition between applicants, common solutions for common challenges, improved horizontal policy coordination, reduced programme management costs etc. Public–public partnerships usually involve the EU making available a suitable legal framework (based on Article 185 of the Treaty) as well as 'seed money' that should incentivise Member States to put together their research funding into a common pot.

The European Research Council awards research grants to individual researchers for excellent frontier research. It was developed based on the observation that Europe is lagging behind in terms of research excellence, as reflected in *inter alia* highly cited publications. The European Research Council constitutes an entirely new kind of FP initiative in the sense that its rationale is based not on the usual cross-border collaboration and coordination arguments traditionally underpinning FP actions, but on an entirely new type of European added value argument: the EU produces added value by letting researchers compete at the pan-European level rather than at the Member State level, thereby raising the intensity of competition, which raises the quality of research proposals and projects and contributes to European excellence in research.

Public–private partnerships promote the development and execution of longerterm strategic research agendas in areas that are critical for Europe's competitiveness, growth and jobs, but which suffer from large-scale market and systemic failures due to the scale of investment needed, the degree of coordination between stakeholders required, the market risk involved etc. These initiatives were developed based on the observation that in areas perceived as critical for Europe's industrial future (e.g. pharmaceuticals, aeronautics, energy) Europe was starting to lag behind in terms of research, innovation and deployment.

The Risk-sharing Finance Facility (RSFF), involving in addition to the European Commission the European Investment Bank, is a facility to finance higher risk research, technological development, demonstration and innovation investments. The rationale for the RSFF is that one of the key factors constraining the implementation of research and innovation activities is the insufficient availability of financing, at acceptable terms, to promoters of investments involving complex products and technologies, unproven markets and intangible assets. In order to overcome these difficulties, the RSFF improves access to debt financing for private companies or public institutions promoting research and innovation activities.

The thematic coverage of the FP has also expanded. While the FPs initially focused on energy and ICT, currently the whole range of scientific and technological fields is covered. The same applies to the range of instruments.

In accordance with the expansion in terms of budget, scope, thematic overage and instruments, the level of participation in the FP has increased over time. In the cross-border collaborative research projects, for instance, the number of participants has increased from 13,000 under FP2 to 40,000 under FP6.

Over time, the FP has become more academic in nature, as industrial participants, initially accounting for the majority of the aforementioned participations, lost out compared to academic institutions. As of FP6, this trend of declining industrial participation has stabilised, however, and even been reversed somewhat. This is due in no small part to the focused attention paid in recent FPs to facilitating the participation by SMEs.

Driving the Evolution of the Framework Programmes: Evaluation and 'Lessons Learned'

The evolution of the Framework Programmes described above has been driven by the lessons learned from numerous interim and *ex post* evaluations. Each FP has been the subject of interim and *ex post* evaluations, which each time have identified weaknesses that subsequent FPs have tried to address. During the preparations of the

most recent FP (Horizon 2020), for instance, it was noted that research, innovation and education should be addressed in a more coordinated manner and coherently with other policies, and that research results should be better disseminated and valorised into new products, processes and services. The intervention logic of EU support programmes should be developed in a more focused, concrete, detailed and transparent manner. Programme access should be improved and start-up, SME, industrial, EU-12 and extra-EU participation increased. Monitoring and evaluation should be strengthened.

A number of FP *ex post* evaluations have noted that the coordination between, on the one hand, the FP and other EU policies, and, on the other hand, the FP and Member State research activities, could be improved. With regard to horizontal policy coordination in the narrow sense, the FP7 interim evaluation (Annerberg et al. 2010) noted that a strategic shift is needed to establish stronger and better connections between research, innovation and education (the so-called knowledge triangle). As for broader horizontal policy coordination, the FP6 ex-post evaluation (Rietschel et al. 2009, pp. 58–59) called for a clearer division of labour between the FP and the cohesion funds. It also stated that other EU policies, such as transportation and energy, would benefit from a more coordinated interface between FP research activities and regulatory and demand-side policies.

With regard to vertical policy coordination, the FP6 *ex post* evaluation noted that, given its small size compared to Member State expenditure, the FP should not try to substitute for Member State R&D policies but should use its added value in a more strategic way and set an attractive and accepted European agenda. In the same vein, European research policy expert Erik Arnold (Arnold 2009, p. 28) concluded that the division of labour between the EU and national levels should be further refined and more explicitly defined, in particular in view of the introduction of the likes of the European Research Council and the Joint Technology Initiatives.

A number of FP *ex-post* evaluations (Rietschel et al. 2009; European Court of Auditors 2007, paragraph IV) have noted that the programme's design could be improved. The view held is that the FP lacks a transparent, clear and robust intervention logic: the programme has too many objectives, and higher-level objectives are insufficiently translated into lower-level objectives. With regard to the FP's objectives, the FP6 *ex post* evaluation (Rietschel et al. 2009, p. vii) as well as expert evidence (Arnold 2005, p. 29) noted that there were too many – addressing almost all science and technology (S&T) and socioeconomic challenges – and that they were too abstract and vague and therefore untestable, complicating *ex post* evaluation. A recent European Parliament ITRE Committee report (ITRE Committee Report 2011, paragraph 9) noted in the same vein that 'an ever-growing number of objectives and themes covered and diversification of instruments has widened the scope of FP7 and reduced its capacity to serve a specific European objective'. In addition, no explicit links are made between higher-level objectives and lower-level concrete technical goals (European Commission 2005, p. 19; Arnold 2009, p. 2). Meanwhile, instruments are not designed explicitly to achieve particular objectives: challenges are defined so as to match existing instruments, not the other way around (Stampfer 2008, p. 13). The result is 'catch all' instruments trying to tackle all problems and to satisfy all types of stakeholders. That is why the European Court of Auditors has called for a system which addresses a single objective in each instrument (European Court of Auditors 2009, paragraph 57.).

All FP *ex post* evaluations – see, for example, the chapters on participation in the FP6 *ex post* (Rietschel et al. 2009) and FP7 interim evaluations (Annerberg et al. 2010) – are unanimous in their view that FP application, contract negotiation and project management procedures are too complex and burdensome, and that this results in high barriers to FP application and participation in general, but for firsttime, start-up, SMEs and EU-12 applicants in particular.

Participants' main reasons for getting involved in the FP relate to networking and the creation of new knowledge (Arnold 2009, p. 2). FP research is also more of a long-term, exploratory, technologically complex nature (Polt et al. 2008). The FP

should not, therefore, be expected to produce new, immediately commercialisable products and processes. Nevertheless, FP evaluations conclude that more attention should be paid to the production of project outputs and to their dissemination and economic valorisation, in particular since the FP is supposed to support Europe's competitiveness. What is highlighted is the absence in the FP of valorisation channels that enable the exploitation of research results and the linking of knowledge created through the FP with socially beneficial uses (Rietschel et al. 2009, pp. 26 and 37; Annerberg et al. 2010, pp. 62 ff). In the same vein, the FP7 interim evaluation observes a lack of clarity on how the FP incorporates innovation (as opposed to 'pure' research).

The main problem affecting the FP monitoring and evaluation system relates to the aforementioned lack of focused objectives and a robust intervention logic. The evaluation process aims to link evidence emerging from project implementation with the strategic and specific objectives set for the programme. As the European Court of Auditors (2007) observed, if this connection is difficult to make, an assessment exercise becomes extremely complicated.

The Impacts of the Framework Programmes

Throughout their existence, the FP have been thoroughly evaluated and they have been demonstrated to have achieved large-scale impacts. Thus it has been shown that the FP have involved large numbers of top (A-team) EU and extra-EU researchers in thousands of first rate, mixed (firms, universities, research institutes), cross-border projects carrying out excellent, often interdisciplinary, collaborative research on a very wide range of topics. For instance, based *inter alia* on an FP-wide bibliometric study that demonstrated that the publication and citation performance of FP project 'lead scientists' is better than that of their non-FP peers (EPEC 2009), the FP6 *ex post* evaluation (Rietschel et al. 2009) concluded that FP6 involved top-quality researchers in first-rate projects performing high-quality research. Observing *inter alia* that 'the list of organisations that have obtained the largest amounts of funding from FP7 can be read as a Who's Who of European research quality', the FP7 interim evaluation (Annerberg et al. 2010) concluded that 'there can be little doubt that FP7 attracts the top EU researchers from universities and RTOs'. According to a Dutch FP impact study (Technopolis 2009), 'bibliometric research and over 100 interviews held in the Netherlands, confirmed that the European research programmes produce high-quality research and attract the best European researchers'.

The FP has facilitated the training and pan-European/extra-European mobility of researchers, enhanced the quality of doctoral training (including through industrial doctorates), added to the research capabilities of participating institutions, and formalised and oriented the R&D and innovation processes of, in particular, small organisations (e.g. SMEs), young organisations (e.g. start-ups) and organisations from new Member States and candidate countries. For instance, the FP6 *ex post* evaluation (Rietschel et al. 2009) noted that FP6 human resources and mobility schemes involved 8,000 organisations and supported some 12,500 fellows. The FP7 interim evaluation (Annerberg et al. 2010) noted that the specific programme 'People' was making a valuable contribution to the development of researcher human capital and that 'the Marie Curie Actions, through their bottom-up approach, have promoted excellence and have had a pronounced structuring effect on the research landscape'. According to an Irish evaluation of FP6, each project produced, on average, 2.3 newly trained/qualified personnel (Forfás 2009). A study of the impact of FP6 in new Member States (COWI 2009) found that FP6 'had an important impact on research organisations' interests and capacity in networking and . . . inspired a networking approach to the management and implementation of research projects with more focus on cooperation, formation of consortia, multidisciplinarity, communication and management skills'. It also produced 'an increase in skills and research capabilities of its key research staff' and resulted in the 'development of administrative capacity/competence to handle international project management processes'.

The FP has produced new knowledge embodied in large numbers of influential (because highly cited) (co-)publications and enhanced the development of new products and processes; the development and use of new tools and techniques; the design and testing of models and simulations; the production of prototypes, demonstrators and pilots; and other forms of technological development. For instance, according to Forfás (2009) each project produced, on average, 12.7 publications (of which 5.3 were in refereed journals and books) and 5.2 conferences, seminars and workshops.

The FP has generated large numbers of patents and enabled participants to increase their turnover and profitability, raise their productivity, increase their market share, obtain access to new markets, reorient their commercial strategy, improve their competitive position, enhance their reputation and image, and reduce commercial risk. For instance, according to an FP6-wide survey (IDEA Consult 2009), industrial organisations clearly expected commercial returns. Almost half of them (47%) stated that these were 'likely' to 'very likely', and 60% of this group expected these returns within 2 years (90% within 5 years). According to the FP5 and FP6 Innovation impact study (Polt et al. 2008), the great majority of FP participants reported at least one form of commercialisable output (new or improved processes, products, services, standards) stemming from their FP project and a large number even recorded more than one such output; an econometric analysis showed that the FP produces output additionality – a positive impact on the innovative sales of firms participating in the FP; and small and medium-sized enterprises indicated the most positive results in terms of innovation in FP projects. According to a German evaluation of FP6 (Federal Ministry of Education and Research 2009), scientific personnel participating in FP6 stated that a substantial part of their patent applications was due to their participation in the FP. According to a UK evaluation of the FP (Technopolis 2010), a majority of UK business participants stated that their involvement in the FP had yielded important commercial benefits; in terms of immediate project outputs, a significant proportion of business respondents reported having made or gained access to new or significantly improved tools or methodologies, and in a large minority of cases firms reported the creation of formal elements of intellectual property; beyond these immediate project results, around 20% of businesses stated that their participation had made significant contributions to the development of new products and processes and in around 10% of cases organisations reported increased income and market share. Lastly, company interviews suggested that FP participation had made a significant contribution to the competitiveness of leading players in several niche technology markets, from inkjets to photonics.

In addition, the results of FP direct and indirect actions have supported EU-level policy formulation. For instance, according to an EC-commissioned evaluation of FP6 environmental research (EPEC 2008), at the international level EU research related to climate change contributed to the International Panel on Climate Change (IPCC), either directly, through individual researchers involved in the IPCC review, or through references to EU-funded projects in IPCC reports.

The FP's positive impacts on innovation have translated, down the line, into large-scale positive macroeconomic, social and environmental impacts. For instance, according to econometric analyses underpinning the Horizon 2020 *ex ante* impact assessment, Horizon 2020 will help to boost industrial productivity, with every €1 invested generating an average of €13 in increased added value of the business sector. Investing €11.5 billion per year on average under Horizon 2020 (and sustaining this investment in the years thereafter) will, by 2030, generate €115 billion per year of extra GDP. This means that by 2030 each euro invested annually in research and innovation under Horizon 2020 will generate around 10 euros of extra GDP. In addition, Horizon 2020 will create about 830,000 durable jobs by 2030 or one job per 100,000 euros disbursed under Horizon 2020.

The FP has produced so-called structuring effects: durable changes in the EU research and innovation landscape. If it were not for the FP, the European Research Council, promoting excellence across Europe, would not have been created; the EU would then have been left with a landscape of compartmentalised national research councils, but

would have had no funding mechanism to promote EU-wide competition for funds and to encourage higher scientific quality in frontier research. As a result of the Marie Curie Actions, the EU has created the right framework for researchers' careers and the free movement of knowledge. The EU leads in the creation and use of research infrastructures of pan-European importance: as a result of EU leadership, for the first time, a pan-European strategy on research infrastructures (the so-called ESFRI roadmap) has been developed and is now being implemented. Collaborative research projects, international cooperation actions, mobility actions, and research infrastructure actions have generated durable, cross-sectoral, interdisciplinary research and innovation networks across Europe as well as with the world's most dynamic and fastest growing research nations that have survived after the end of EU funding. European Technology Platforms and ERA-NETs have served as useful focusing devices that have helped stakeholders identify and explain their R&D needs jointly, easing the process of developing mutually supportive policies at EU and Member State levels. Joint technology initiatives have focused and aligned key actors in their respective areas, serving as a support to develop coherent sectorial strategies. Article 185 and joint programming initiatives have achieved a better coordination of R&D in Europe and supported a more coherent use of resources.

It is important to emphasise that the evaluation literature has convincingly demonstrated that, in the absence of EU funding, these projects would not have been carried out, or would have been postponed or scaled down in financial terms, in terms of scope and ambition, or in terms of the number of partners involved. In other words, the FPs achieve large-scale additionality effects.

The Increasing Importance of Policy Advocacy

While the main policy tool for implementing the EU's R&D policy so far has been direct financial support disbursed through the aforementioned multi-annual Framework Programme, policy advocacy has gained in importance over the past 10–15 years. This started with encouraging the Member States to invest 3% of their GDP in R&D within the context of the Lisbon Strategy. This still continues with the 'soft' approach being taken towards the achievement of a European Research Area, which is defined as 'a unified research area open to the world based on the Internal Market, in which researchers, scientific knowledge and technology circulate freely'. And this has taken on a larger scale still with the so-called 'Innovation Union' flagship initiative of 'Europe 2020', which is the European Union strategy to create an innovation-friendly environment that makes it easier for ideas to be turned into products and services that will bring the economy growth and jobs and which pursues improved access to finance, innovation-friendly rules and regulations, accelerated standard-setting, cheaper patenting, innovation supported by the public sector, innovation partnerships to give EU businesses a competitive edge, facilitated access to EU research and innovation programmes.

See Also

- ▶ European Union (EU) Trade Policy
- ▶ Research and experimental development (R&D) and technological innovation policy
- ▶ Research Joint Ventures
- ▶ European Union's Common Agricultural Policy (CAP)
- ▶ Theory of Economic Integration: A Review

Appendix: European Added Value – Why Fund Research at EU Level?

EU support to research is provided only when it can be more effective than national funding. It does this through measures to coordinate national funding, and through implementing collaborative research and mobility actions.

Coordinated Funding and Agenda-Setting

EU initiatives help to coordinate funding across national borders and to restructure the R&D and innovation landscape in Europe:

- The EU has created the European Research Council. Without it, the EU would have a landscape of compartmentalised national research councils, but no mechanism to promote EU-wide competition for funds and to encourage higher scientific quality.
- Thanks to EU leadership, for the first time, a pan-European strategy on research infrastructures is now being implemented.
- The EU helps private companies come together and implement joint strategic research agendas through tailored instruments, such as European Technology Platforms and Joint Technology Initiatives.
- The EU joins up compartmentalised national research funding using instruments such as ERA-NETs and Article 185 initiatives, which set common agendas and achieve the funding scale required for tackling important societal challenges.
- Through its Marie Curie actions, the EU set standards for innovative research training and career development and put in place a framework for the free movement of knowledge.

Coordinated funding reduces duplication and increases efficiency. EU support is vital – none of the above measures would have seen the light of day without an EU initiative.

Collaborative Research Projects and Mobility Actions

When it comes to implementing research and innovation projects, EU actions add value by stimulating transnational collaboration and mobility. These actions generate a series of benefits that could not be achieved by Member States acting alone:

- Support for collaboration helps to achieve the critical mass required for breakthroughs when research activities are of such a scale and complexity that no single Member State can provide the necessary resources.
- The EU supports research which addresses pan-European policy challenges (e.g. environment, health, food safety, climate change, security), and facilitates the establishment of a common scientific base and of harmonized laws in these areas.
- Working in trans-national consortia helps firms to lower research risks, enabling certain research to take place. Involving key EU industry players and end users reduces commercial risks by aiding the development of standards and interoperable solutions, and by defragmenting existing markets.
- Collaborative research projects involving end users enable the rapid and wide dissemination of results leading to better exploitation and a larger impact than would be possible only at Member State level.
- SME involvement in research and innovation at EU level improves their partnerships with other companies and labs across Europe, and enables them to tap into Europe's creative and innovative skills potential, to develop new products and services, and to enter new national, EU or international markets.
- Companies can collaborate with foreign partners and end users at a scale not possible at national level, in projects tested for excellence and market impact, which induces them to invest more of their own funds than they would under national schemes.
- Cross-border mobility and training actions are of critical importance for providing access to complementary knowledge, attracting young people into research, encouraging top researchers to come to Europe, ensuring excellent skills for future generations of scientists, and improving career prospects for researchers in both public and private sectors.

Bibliography

Annerberg, R., I. Begg, H. Acheson, S. Borrás, A. Hallén, T. Maimets, R. Mustonen, H. Raffler, J.P. Swings, and K. Ylihonko. 2010. *Interim evaluation of the seventh framework programme – Report of the expert group*. European Commission, Directorate-General for Research.

Arnold, E. 2005. *What the evaluation record tells us about framework programme performance*. Brighton: Technopolis Group.

Arnold, E. 2009. *Framework programme 6 – Meta-evaluation*. Brighton: Technopolis Group.
COWI. 2009. *Assessment of the impact of the 6th framework programme on new member States*. European Commission, Directorate-General for Research.
Dahrendorf, R. 1973. *Towards a European science policy.* Southampton: University of Southampton.
EPEC. 2008. *Ex-post Impact Assessment FP6 Subpriority* '*Global change and ecosystems*'. Final Report, European Policy Evaluation Consortium (EPEC) and Technopolis Ltd for the Directorate-General for Research.
EPEC. 2009. *Bibliometric profiling of framework programme participants – Final report*. European Policy Evaluation Consortium (EPEC), European Commission, Directorate-General for Research.
European Commission. 1997. *EU research and technological development activities, 5-year assessment of the European community RTD framework programmes: Report of the independent expert panel – Commission's comments – Communication from the commission, COM(97) 151 final*. Brussels: European Commission.
European Commission. 2005. *Five-year assessment of the European union research framework programmes 1999–2003*. Luxembourg: Office for Official Publications of the European Communities.
European Commission. 2013. *The grand challenge. The design and societal impact of horizon 2020*. Luxembourg: Publications Office of the European Union.
European Court of Auditors. 2007. *Evaluating the EU RTD FP – Could the commission's approach be improved?* Special Report no 9/2007.
European Court of Auditors. 2009. '*Networks of excellence*' *and* '*Integrated projects*' *in community research policy: Did they achieve their objectives?* Special Report no 8/2009.
Federal Ministry of Education and Research. 2009. *German participation in the sixth European framework programme for research and technological development*. Berlin: Bonn.
Forfás. 2009. *Evaluation of framework programme 6 in Ireland*. Forfás.
IDEA Consult. 2009. *Participation survey and assessment of the impact of the actions completed under the 6th framework programme*. European Commission, Directorate-General for Research.
ITRE Committee Report. 2011. *Report on FP7 interim evaluation*, 2011/2043(INI), rapporteur: J. P. Audy. European Parliament.
Majó, J., G. Argyropoulos, S. Barabaschi, J. Bell, H. Danielmeyer, Y. Farge, S. McKenna-Lawlor, F. Thys-Clement, C. Ullenius, J. Viana Baptista, N. Wilhjelm, and K. Guy. 2000. *Five-year assessment of the European union research and technological development programmes, 1995–1999 – Report of the independent expert panel*. Brussels: European Commission.
Muldur, U., F. Corvers, H. Delanghe, J. Dratwa, D. Heimberger, B. Sloan, and S. Vanslembrouck. 2006. *A new deal for an effective European research policy. The design and impacts of the 7th framework programme*. Dordrecht: Springer.
Polt, W., N. Vonortas, and R. Fisher. 2008. *Innovation impact study – Final report*. Brussels: DG Research.
Rietschel, E., E. Arnold, C. Antanas, A. Dearing, I Feller, S. Joussaume, A. Kaloudis, L. Lange, J. Langer, V. Ley, R. Mustonen, D. Pooley, and N. Stame. 2009. *Evaluation of the sixth framework programmes for research and technological development 2002–2006 – Report of the expert group*. European Commission.
Stampfer, M. 2008. *European added value of community research activities – Expert analysis in support of the Ex post evaluation of FP6*. WWTF – Vienna Science and Technology Fund.
Technopolis. 2009. *Impact Europese kaderprogramma's in Nederland*. Brighton: Technopolis Group.
Technopolis. 2010. *The impact of the EU RTD framework programme on the UK*. Office of Science and Technology.

European Union (EU) Trade Policy

Stephen Woolcock

Abstract

The European Union's (EU) role in international trade has evolved from a defensive position during the 1960s and 1970s, to being a firm supporter of a rule-based multilateral trading system as a member of the Quad (US, EU, Japan and Canada) in the 1980s and to a role in which it aspires to leadership. Shifts in relative market power with the rise of emerging markets has, however, undermined the EU's ability to shape outcomes. Thanks to a well-developed internal *acquis*, the EU has developed common policies on all trade and trade-related topics, but the normative power this provides has had little discernable impact on multilateral trade outcomes. The decision-making procedures of the EU have functioned tolerably well up to now thanks to Member States having confidence and trust in the way decisions are made and the way the Commission, as agent, is controlled. The need to integrate the European Parliament (EP) into

decision-making procedures following the Lisbon (TFEU) Treaty is, however, likely to result in a period of uncertainty.

Keywords
European community; European Union; GATT; International trade; Multilateralism; WTO

JEL Classifications
F10; F13; F20

Introduction

This contribution provides an overview of the evolution of EU policy, a summary of the EU's positions on key issues in international trade and a summary of the decision-making procedures in EU external trade policy after the adoption of the Lisbon Treaty. The article therefore provides an introduction to the topic as well as sufficient references to enable readers to follow up the various aspects of the topic.

The Evolution of EU Trade Policy

The Treaty of Rome granted the European Economic Community (EEC) exclusive competence for Common Commercial Policy (CCP). (The term 'EU trade policy' will be used as it is the current usage. In actual fact, Common Commercial Policy, the term used in the original Treaty of Rome, more accurately reflects EU policy, which today extends well beyond what has been traditionally considered to be trade policy.) The creation of a customs union required the adoption of a common external tariff and thus a single EEC position on tariffs. The customs union also created a collective market power that exceeded that of the individual Member States. As a result the EEC was able to achieve some important offensive interests during the Kennedy Round (1963–1966) of the General Agreement on Tariffs and Trade (GATT), notably a reduction in US tariffs (Duer 2008). A desire to show solidarity in building Europe and a decision-making process that enabled Member States to veto trade concessions also enabled the EEC to hold its defensive positions. These were to retain the preference margin for EEC producers that the customs union would create and to protect the fledgling Common Agricultural Policy (CAP).

In the 1970s the USA again led the charge in GATT. Facing a deteriorating balance of trade and what it saw as 'unfair' trade practices of the Japanese and Europeans in supporting their national industries, the USA pushed for multilateral controls for subsidies, an opening of government procurement markets and disciplines covering technical regulations and standards. The USA had no active industrial policy, and had decentralized public purchasing and standards setting, so it viewed the coordination of such instruments to favour national companies in other countries as unfair. But European Community (EC) Member States pursued explicit (France and Britain) or implicit (Federal Republic of Germany) national champion strategies. The implications for EC trade policy were, however, the same: namely the defence of the policy space to enable these national policies to be continued.

There was some debate on EC-level industrial policy, but Member States' interests were too divergent for such an active policy. The only EC level intervention was in the form of coordinated adjustment or restructuring (in the face of competition from Japanese and Asian Newly Industrializing Countries) (Turner et al. 1982). Towards the end of the 1970s there was some support for what was called pre-competitive cooperation between producers in different Member States in more advanced technology sectors (McGuire 2006). This was not significant except that it heralded a shift in private sector opinion away from reliance on national markets and towards greater market integration within Europe in order to compete with Japan and the USA.

The EC also entered the 1980s with a defensive position on international trade and resisted new initiatives on non-tariff barriers, services, investment and intellectual property rights (IPRs). But a paradigm shift within the EC towards more

liberal, rule-based policies facilitated both the Single European Market (SEM) initiative and thus support for a more proactive EC position on international trade (Young and Peterson 2006). The SEM embodied a compromise between French dirigisme and rigorous reciprocity in trade negotiations and Anglo-Saxon liberalism. In fact, the outcome was closer to a form of EC 'Ordungspolitik': in other words a policy based on competition within the market and within an agreed framework of regulations guaranteeing key non-economic objectives and competition (Hodges et al. 1991). The SEM and associated introduction of qualified majority voting with the Single European Act (SEA) resulted in common EC approaches to almost all the issues under discussion in the GATT Uruguay Round between 1986 and 1994. There were effective EC level controls of subsidies, a comprehensive regime established for government procurement covering all forms of contract and levels of government. Technical regulations were addressed by the 'New Approach', a combination of harmonization of minimum essential requirements and mutual recognition. The SEM also liberalized some key service sectors, including financial services and telecommunications, two sectors which were seen as priorities for multilateral liberalization.

The deepening and widening of the SEM enhanced EU market power (Holmes 2006). The strengthened *acquis communautaire* gave the EU 'normative power', as did consensus on the balance between market and regulation. Acceptance of a liberal, rules-based regime within the EU meant that the EU was ready to support an equivalent regime at multilateral level provided it was consistent with the EU rules. Taken together, these factors enabled the EU to play an active role in the Uruguay Round, and the EU together with the USA (and other members of the Quad) shaped the agenda and very largely the outcome of trade negotiations. On more traditional trade issues the EU further reduced its bound tariffs on manufactures to an average of about 4 per cent. On agriculture the EU fought a rearguard action against liberalization and in the end accepted the reestablishment of multilateral rules for agriculture, but little in terms of actual liberalization.

From the mid-1990s the EU became the main proponent of a new multilateral round of trade negotiations. With the USA reluctant to engage in further multilateral liberalization due to domestic opposition and developing countries largely opposed to a comprehensive round, the EU assumed a kind of leadership role. The EU approach was shaped by the European Commission, which favoured a new comprehensive multilateral round ahead of preferential trade agreements (Lamy 2004). For the EU a comprehensive round meant coverage of the issues already covered by the GATT/WTO, such as tariffs, non-tariff barriers, services and agriculture, as well as issues for which there were as yet no established multilateral rules, such as investment, competition policy, government procurement and trade facilitation. These four issues became known as the 'Singapore issues' because the EU ensured they were placed on the WTO work programme at the WTO ministerial meeting in Singapore in 1996. Trade and labour standards and trade and the environment, or 'sustainable trade' as the coverage of these two topics have come to be known in the EU, were also discussed at the Singapore ministerial. The EU did not push hard for these topics to be added to the WTO work programme, because opinion within the EU was divided between Member States such as France and some socially minded northern Member States that favoured the inclusion of labour standards, and others, such as Britain and Germany, that opposed discussing labour standards in the WTO. With developing country members of the WTO firmly opposed to including environment and labour standards in talks, the European Commission did not press the issue.

By the time a multilateral round, in the shape of the Doha Development Agenda (DDA), was launched, some sources of EU relative strength had been weakened (Young and Peterson 2006). The Lisbon agenda, a more intergovernmental follow up to the SEM, proved largely unsuccessful. As the 2000s progressed, growth in relatively high tariff and otherwise protected markets such as China, India and Brazil burgeoned, eroding the EU's relative market power.

The negotiating leverage gained from holding out the prospect of concessions on agriculture in

order to pursue offensive interests in non-agricultural market access (NAMA), services and the Singapore issues proved insufficient, and at the WTO Ministerial meeting in Cancun in 2003 the EU was obliged to drop investment, competition and government procurement from the agenda. At the Hong Kong WTO Ministerial meeting in 2005 the EU made further concessions, such as agreeing to phase out export subsidies in agriculture, in order to keep the round alive. But with little support from the USA, and opposition from developing countries, the aim of an ambitious comprehensive round was lost and the DDA reverted to a modest conventional market access trade round focused on agriculture and NAMA. By 2006 the EU recognized the that success at the multilateral level was unlikely and switched to bilateral negotiations with major potential markets, especially in Asia, as well as complete existing negotiations with African Caribbean and Pacific (ACP) states (Elsig 2007; Heydon and Woolcock 2009; Bartels 2007). This policy was codified in the October 2006 Global Europe Strategy (European Commission 2006; Evenett 2007) and has been subsequently confirmed by the November 2010 policy statement on Trade, Growth and World Affairs (European Commission 2010). The 2010 policy statement also suggests a less liberal approach to reciprocity by hinting at the potential withdrawal of access to procurement markets for trading partners that do not offer reciprocal access.

A Summary of EU Policy Positions

In addition to the kind of developments in European integration and the international trading system discussed above, EU trade policy is shaped by sector interests. Indeed, the general structure of EU preferences can be traced to the balance between offensive and defensive interests of sectors of the Member State economies. Such sector interests are of course aggregated in EU level policies, so that the *acquis* itself reflects the balance of sector preferences.

The EU has generally favoured a formula approach to tariff reductions because the creation of the common external tariff (CET) appears to have smoothed the EU's tariff profile so that it has had rather higher average tariffs but fewer tariff peaks than, for example, the USA. Consecutive multilateral rounds have reduced the average MFN tariff for manufactured goods to 3.9 per cent with 100 per cent tariff binding. The EU therefore has less to offer in NAMA compared to the large emerging markets such as China, Brazil and India (9 per cent, 12 per cent and 16 per cent applied rates respectively) with higher bound rates in some cases. (The bound rate is the rate bound under GATT commitments. A higher bound rate than an applied rate means that a WTO member can increase tariffs up to the bound rate without infringing GATT rules and thus facing retaliation from other WTO members.) The EU policy position favours significant reductions in bound rates for major emerging markets. Such reductions in bound rates will not result in any significant reductions of applied rates, so what the EU seeks is discipline to prevent the emerging markets increasing rates on EU exports thanks to 'water' in their tariffs or relatively high bound rates. For least developed countries the EU has offered tariff free quota free access to the EU market and urges other major WTO members to do the same (Faber and Orbie 2007). The EU also supports sector negotiations, there are such negotiations in 14 sectors, and seeks some commitment from the large emerging markets to this process.

The average EU tariff is 15 per cent in agriculture, compared to 10 per cent in Brazil and China and 38 per cent in India, but the EU of course provides significant agricultural subsidies. Since the initial limited McSharry reforms of 1992, the EU further reduced price support levels and 'decoupled' agricultural support from trade in the Agenda 2000 reform and especially the mid term review of the CAP in July 2003 (Daugbjerg and Swinbank 2009). These provided some scope for the EU to make concessions so that it has accepted tiered tariff and subsidy reductions in the chair's text of December 2008 in the DDA. (The tiers as set out in the Chair's text of December 2008 are tariffs of more than 75 per cent (70 per cent reduction), 50–75 per cent (64 per cent reduction), 20–50 per cent (57 per cent) and

less than 20 per cent (54 per cent reduction).) If finally agreed, these could lead to further market opening by the EU, but much depends on the detail, including in particular what percentage of product lines are defined as sensitive and therefore excluded from tariff reductions. Anything more than 4 per cent would limit liberalization considerably. On subsidies the Chair's text would result in an 80 per cent reduction in of the Overall Trade Distorting Support (OTDS) for EU agricultural, with some safeguards against shifting of subsidies between activities. This represents further liberalization, although it is of course dependent on an agreed outcome of the DDA round as a whole. As part of a strategy of diversification out of commodity crops and into higher value-added agricultural products, the EU is seeking greater protection for geographic indications, such as Parma ham and champagne.

The EU has pushed for the Singapore issues. On public procurement, the EU, having adopted a comprehensive EU regime internally, would like other major economies that have not signed the WTO's Government Purchasing Agreement (GPA) to at least adopt measures on transparency. The EU policy also believes that greater transparency is beneficial for all countries because it promotes competition and more efficient use of public finance, and fights corruption in the allocation of public contracts.

The EU supported negotiating investment in the WTO, in part to include investment in the rules-based regime of the WTO and in part because the main restrictions on investment were in developing countries. The USA favoured the plurilateral OECD because it wanted higher standards than could be expected in any WTO agreement. The collapse of the plurilateral 'Multilateral' Agreement on Investment (MAI) in 1998 also effectively ended prospects for agreement within the WTO. The recent extension of the EU's exclusive competence to foreign direct investment with the Lisbon Treaty can be expected, in time, to result in a more common approach to investment by the EU (European Parliament 2010). In the past the EU's policy has been hampered by the fact that competence for foreign direct investment was shared between the European Union and the Member States, with the Member States leading in negotiating investment protection in bilateral investment treaties (BITs). Exclusive competence implies the need to define a comprehensive, common EU position on investment. Given the importance of the EU for foreign direct investment a redefined, 'modern' approach to investment agreements by the EU could breathe some life into the prospects of a genuine international agreement.

The rationale for EU support for the inclusion of competition as one of the Singapore issues was that there was a need to ensure that private restraints to trade do not replace public constraints following liberalization. This was the same rationale used for EU-wide competition policy. Despite difficulties gathering information, there is evidence of damaging international cartel activity. The European Commission led in pushing for the inclusion of competition because it has exclusive powers in this policy area (Damro 2006). But there was little support except among consumer groups within the EU. Internationally there was strong opposition from the USA, where the Department of Justice opposed any substantive international rules on competition, and resistance from developing countries, which argued that they did not have the capacity for such policies.

In services the EU retains an offensive position given its comparative advantage in many service sectors, such as financial services and business services. Since the financial crisis of 2008 the mood has swung against further liberalization of financial services. In the field of intellectual property rights the EU favours more effective enforcement of existing international conventions. Finally, with regard to technical regulations and sanitary and phytosanitary measures the EU appears to have shifted from a policy of seeking mutual recognition agreements, because of the complexities involved in these, and now favours the promotion of full use of existing international standards.

The Policy Process

EU policy-making in trade functions reasonably well, despite the need to reconcile the positions of 27 Member States, when there is a strong internal

consensus as with the SEM and when there is a well-established decision making regime in which the major stakeholders have confidence (De Bievre and Duerr 2007). The decision-making regime for external trade has been established over a period of 50 years since the Treaty of Rome and has provided the model for all external EU policy-making used in the Treaty of Lisbon (Art 218 TFEU). In this regime the Commission provides the strategic orientation of policy thanks to its right of initiative on negotiating mandates. The EU's negotiating aims or mandate are adopted by the Member States in the Foreign Affairs Council after work in the Trade Policy Committee (ex Art 133 Committee) that brings together Member State and Commission senior trade officials (Art 207(3) and 218(2) TFEU). The TFEU confirmed that the consent of the EP, by a simple majority, is needed for all trade and investment agreement negotiated by the Commission. The EP would also like more say in setting EU objectives, because making its consent to any trade agreements conditional upon certain targets being met would strengthen the credibility of the veto power. The TFEU does not provide for this.

During negotiations, whether at multilateral or bilateral levels, at a technical or political level, the European Commission is the sole voice of the EU (Young 2006). This greatly facilitates coordination compared to other policy areas, where there are different negotiators at the technical and political levels, such as in the case of international environmental policy. There remain of course coordination problems both within the Commission between Directorates General and between the Commission and the Member States (Kerremans 2006; Meunier and Nicolaidis 2006). The well-established regime of decision-making in which the Member States, through the TPC, assist the Commission during negotiations has generally promoted trust between Commission and Council. Such close supervision of the Commission's approach to negotiations provides the assurance needed by the Member State governments to allow the Commission to negotiate. The Council can also give the Commission directions during negotiations.

The adoption of the TFEU now requires the European Parliament must now be included in decision-making including during negotiations. The EP (International Trade Committee (INTA)) now receives the same information on the progress of negotiations as the TPC. The Commission has for some time been working more closely with the EP in anticipation of the treaty changes and has already provided a great deal of information (Woolcock 2010). The Council and Member States have a less easy relationship with the EP on trade policy and it will take some time before a *modus vivendi* can be developed between the two. Both Commission and Council, as well of course as interest groups and lobbies, will have to pay more attention to the EP, which has power to grant consent to all trade agreements. The EP also shares powers with the Council on trade legislation, such as the adoption of EU legislation implementing trade agreements or so-called autonomous trade measures, such as the Generalized System of Preferences for developing countries. Prior to the TFEU the Council used to adopt legislation according to the coordination procedure in which the EP played virtually no role. After the TFEU the Ordinary Legislative Procedure (OLP) (formerly co-decision making) is to be used. This will be much slower than the previous arrangements, so it is likely that the Commission will be granted implementing powers to deal with the numerous detailed adjustments needed to trade agreements and schedules, with OLP used only for the relatively few major pieces of trade legislation.

Conclusions

EU trade policy has gone through various stages, some more defensive than others. During the 1980s and 1990s the EU moved to become more supportive of a liberal rules-based multilateral trading order. EU efforts to lead a comprehensive WTO round, in the shape of the DDA during the 2000s, has not had much success. As a result the EU has reverted to pursuing bilateral free trade agreements in order to pursue its aims.

The EU policy stance remains generally liberal, with the exception of agriculture, where

reform has been steady but slow, and there is unlikely to be support among a qualified majority of Member States for any move towards the more aggressive use of reciprocity by threatening to close the EU market. But the EU lacks much leverage in negotiations, especially multilateral negotiations, due to the fact that it has an open market in most sectors, again with the exception of agriculture. The negotiation coinage that could be offered by way of opening the EU agricultural market did not prove sufficient to make progress on the EU's offensive interests in the DDA.

The decision-making procedures of the EU have functioned tolerably well up to now thanks to Member States having confidence and trust in the way decisions are made and the way the Commission, as agent, is controlled. The need to integrate the EP into the decision-making procedures following the Lisbon (TFEU) Treaty is, however, likely to result in a period of uncertainty.

See Also

- ▶ European Union's Common Agricultural Policy (CAP)
- ▶ International Trade
- ▶ International Trade Theory

Bibliography

Bartels, L. 2007. The trade and development policy of the European Union. *European Journal of International Law* 18(4): 715–756.

Damro, C. 2006. The new trade politics and EU competition policy: Shopping for convergence and co-operation. *Journal of European Public Policy* 13(6): 867–886.

Daugbjerg, C., and A. Swinbank. 2009. *Ideas, institutions, and trade: The WTO and the curious role of EU farm policy in trade liberalization*. Oxford: Oxford University Press.

De Bièvre, D., and A. Duerr. 2007. Interest group influence on policymaking in Europe and the United States. *Special Issue of the Journal of Public Policy* 27(1). Cambridge University Press (February).

Duer, A. 2008. Bargaining power and trade liberalization: European external trade policies in the 1960s. *European Journal of International Relations* 4: 645–671.

Dür, A., and H. Zimmermann, eds. 2007. Introduction: The EU in international trade negotiations. *Journal of Common Market Studies* 45(4): 771–787.

Elsig, M. 2007. The EU's choice of regulatory venues for trade negotiations: A tale of agency power? *Journal of Common Market Studies* 45(4): 927–948.

European Commission. 2006. *Global Europe: competing in the world. A contribution to the EU's growth and jobs strategy*. COM(2006) 567 final.

European Commission. 2010. *Trade growth and world affairs: Trade policy as a core component of the EU's 2020 strategy*. COM (2010) 612 October 2010.

European Parliament. 2010. *Directorate-General for external policies*. *The EU approach to international investment policy after the Lisbon treaty*. INTA PE 433.854-855-856.

Evenett, S. 2007. *'Global Europe' an initial assessment of the European commission's new trade policy*. http://www.evenett.com/articles.htm

Faber, G., and J. Orbie, eds. 2007. *European Union trade politics and development: Everything but arms unravelled*. London: Routledge.

Heydon, K., and S. Woolcock. 2009. *The rise of bilateralism: Comparing American, European and Asian approaches to preferential trade agreements*. Tokyo: United Nations University Press.

Hodges, M., K. Schreiber, and S. Woolcock. 1991. *Britain, Germany and 1992: The limits of deregulation*. London: RIIA.

Holmes, P. 2006. Trade and 'domestic' policies: The European mix. *Journal of European Public Policy* 13(6): 815–831.

Kerremans, B. 2006. Proactive policy entrepreneur or risk minimizer? A principal-agent interpretation of the EU's role in the WTO. In *The European Union's Roles in International Politics*, ed. O. Elgström and M. Smith. Oxford: Routledge.

Lamy, P. 2004. *Trade policy in the prodi commission, 1999–2004*. European commission. An assessment. http://trade.ec.europa.eu/doclib/docs/2006/september/tradoc_120087.pdf

McGuire, S. 2006. No more Euro-champions? The interaction of EU industrial and trade policies. *Journal of European Public Policy* 13(6): 887–905.

Meunier, S. 2000. What single voice? European institutions and EU–U.S. trade negotiations. *International Organization* 54(1): 103–135.

Meunier, S., and K. Nicolaidis. 2006. The European Union as a conflicted trade power. *Journal of European Public Policy* 13(6): 906–925.

Turner, L., N. McMullen, and S. Woolcock. 1982. The newly industrialising countries: Trade and adjustment policies, with, George Allen and Unwin.

Woolcock, S. 2003. *The Singapore issues in Cancun: A failed negotiation ploy or a litmus test for global governance*. Intereconomics, December.

Woolcock, S. 2010. *The treaty of Lisbon and the European Union as an actor in international trade*. ECIPE working paper *n*o 1/2010, January.

Young, A.R., and J. Peterson. 2006. The EU and the new trade politics. *Journal of European Public Policy* 13(6): 795–814.

European Union Budget

Mojmir Mrak

Abstract

The EU budget is a tool through which money is collected and allocated for EU policies and objectives as well as for the tasks transferred to it from the national level. This article starts by presenting the concept and evolution of the EU budget. It then discusses the principles and procedures governing its adoption and implementation, presenting the key features of the expenditure and revenue sides of the EU budget. The main deficiencies of the EU budget are addressed, including the problems of correction mechanisms and net balances. Finally, the challenges faced by the EU at the outset of the negotiations about the post-2013 EU budget are outlined.

Keywords

EU; EU annual budget; EU medium-term financial perspectives; Public finances

JEL Classifications

E42; E5; E50; E52; G1

Introduction

The process of European integration requires financial resources for its activities, and the EU budget is a tool through which money is collected and allocated for EU policies and objectives as well as for the tasks transferred to it from the national level.

The EU budget is modest in size. As agreed by the Member States, in the Own Resources Decision (ORD), the maximum ceiling of the EU budget financing is set at 1.24 per cent of the EU GNI (or 1.27 per cent of EU GNP). In practice, however, the EU budget has always remained well below that ceiling. As public finances of the EU Member states are typically between 40 and 45 per cent of their respective GNI, the EU budget is equivalent to just over 2 per cent of the total public finances of the Member States. The EU budget does not represent a significant factor in almost any consolidated national public finance category. Three key segments of public finance expenditures in practically any country – defence, security and public order expenditure – as well as healthcare, are not even included in the EU budget, while the presence of certain other expenditure items, such as education and housing, is minimal. There is another fundamental characteristic which distinguishes the EU budget from national public finances. In contrast to national public finances, which can run deficits, the EU budget is legally required to be in balance each year.

Even though the EU budget is small in size, it is of a tremendous political importance for the overall EU integration process. This can be illustrated by the highly complex procedure that is required for the adoption and implementation of the annual budget, which involves practically all important EU institutions, as well as by the strongly politicized multi-annual financial framework (MAFF) negotiations that set the ceiling costs on major expenditure items of the EU budget over a five to seven year period.

Evolution of the EU Budget

The evolution of the EU budget can be roughly classified into two periods: the first, between 1951 and 1987, was characterised by a move towards the unification of budgetary instruments and the crisis of Community finances in the 1980s; the second, from 1988 until today, has been characterised by features introduced by the 1988 EU budgetary reform.

1951–1987

The public finance system of the EC began to develop in the early 1950s, when in 1951 the

European Coal and Steel Community (ECSC) Treaty was signed. It was followed by the 1957 European Atomic Energy Community (EURATOM) and European Economic Community (EEC) Treaties. Each of these treaties envisaged different budgets for a particular Community, which led to the co-existence of budgets. The ECSC Treaty provided for two budgets – an administrative and an operating budget. The EURATOM Treaty also set up two budgets: an administrative budget and a research and investment budget. The EEC Treaty, on the other hand, established only one, a so-called 'single budget'.

The 1965 merger treaty incorporated the ECSC and EURATOM administrative budgets into the EEC budget, and five years later, in 1970, the Luxembourg Treaty incorporated the EURATOM research and investment budget into the general budget. The outcome of these developments was the formation of two budgets – the general budget and the ECSC operating budget – and the system was in place until 2002. The ECSC Treaty expired in 2002 and therefore the ECSC operating budget ceased to exist. Since then the EU has operated with a single EU budget.

During the first 20 years of the Community's financial system, there were two important developments for the integration of budgetary instruments. The first was the development of common policies. The most notable events, with considerable financial consequences, were probably the creation of the European Agricultural Guidance and Guarantee Fund (EAGGF) in 1962 as an instrument for implementation of the common agricultural policy (CAP), as well as the establishment of two funds for implementation of the cohesion policy: the European Social Fund (ESF) in 1971 and the European Regional Development Fund (ERDF) in 1975. These two policies still constitute about 80 per cent of current EU budget expenditures. The second development was that the initial system, through which the budgets of all the three Communities were financed through a special system of contributions by the Member States, soon proved to be insufficient and unsatisfactory. The need for a better and more efficient system, which would provide sufficient resources, gradually led to a reform of budget financing. Through the 1970 Luxembourg Treaty, a system of so-called 'own resources' was introduced.

The processes of unification of the budgetary instruments, development of common policies and progress towards financial autonomy were inevitably connected with difficult negotiations. The majority of disagreements had been associated with responsibilities and powers that each of the institutions had in budgetary matters. Although decisions in budgetary matters were in theory primarily the exclusive prerogative of the Council, in practice other institutions were involved at various stages of the budgetary procedure. The 1970 Luxembourg Treaty partly formalised such a practice by giving more power to the Parliament. Since the 1975 Brussels Treaty, powers on budgetary matters have been shared between the Council and the Parliament.

The legal, political and institutional structure for governing the Community's finances established in the early 1970s soon proved to be unsustainable over a longer period of time. Relations between Member States, as well as among the European institutions involved in the budgetary adoption procedure, gradually worsened and finally turned into open conflict. An increasing number of incidents made adoption and management of a budget almost impossible. Between 1980 and 1988, approval of four annual budgets was delayed long enough that provisional arrangements in the form of so-called 'twelfths' had to be applied for several months.

Since 1988

The 1986 enlargement to include Spain and Portugal and the conclusion of the Single Act injected new optimism into the Community and provided a sound political base for a thorough reform of the Community's financial system. In 1987, the Commission presented comprehensive reform proposals and in the following year the European Council adopted the broad lines of these proposals. The main political orientation and operational features of each of these orientations were as follows:

The Community should be given sufficient resources to enable it to operate properly. In operational terms, this meant a revision of the ORD whereby a new (fourth) resource was introduced based on Member States' GNP. From that period on, GNP and later on GNI source, represents the balancing item, i.e. it provides the necessary funding for the Community and, from 1992, for the EU budget.

Strict supervision of the expenditures financed by these additional resources should be exercised. This orientation, aimed at making an effective brake on rising agricultural expenditures, was operationalized through an inter-institutional arrangement for budgetary discipline in procedures, of which a multi-annual financial framework instrument was an integral part.

The political orientation of linking the budget contributions of Member States more closely to their levels of relative prosperity was operationalised through a significant reform of the cohesion policy and especially the instrument for its implementation.

Since 1988, the Community/EU budgetary system has remained more or less unchanged in terms of its size expressed as percentage of EU GNP/GNI as well as in terms of the magnitude of its expenditure and its distribution, with agriculture and cohesion spending consuming the majority of the GNI source.

The EU budgetary system continues to be based on two major elements. *First*, the strategic course of the EU public finances and financial framework for the medium-term period is determined in a multi-annual financial perspective (MAFF). The MAFF is basically an agreement among the institutions on budgetary priorities facilitating the budgetary procedure and the management of various programmes. The MAFF allows financial predictability in the development of EU expenditure. Within the framework of the MAFF, the maximum volume and the composition of the foreseeable EU expenditure are indicated. The MAFF fixes the ceilings for particular expenditure headings as well as for the budget as a whole; the cap on spending levels must be set below the own resources ceiling. The MAFF is a product of an inter-institutional agreement between the Commission, Council and Parliament. Although it is not a multi-annual budget, and the annual budgetary procedure remains necessary to decide the next year's budget, the MAFF is not just indicative, as it sets the maximum ceilings for each year and each category of expenditure (heading). Until now, Community/EU institutions have adopted four MAFFs. The first, called 'Delors I', had duration of five years (1988–1992) while all three of the subsequent ones covered seven-year periods: 'Delors II' (1993–1999), 'Agenda 2000' (2000–2006) and the current MAFF (2007–2013). *Second*, the implementation and operational details of the EU budgetary system are elaborated in the annual budget, which must be consistent with the MAFF.

Principles and Procedures for the Annual Budget and for the MAFF Principles

The EU budget is regulated by six principles that are enshrined either in the Treaty or in the secondary financial legislation.

The *principle of unity* states that all expenditures and revenues of the EU must be included in the EU budget. The European development fund (international development aid) and the financial activities of the European Investment Bank constitute exceptions to this principle.

The *principle of universality* says that revenues cannot be appropriated for specific spending purposes.

The *principle of annuality* requires that budgetary appropriations must refer to a specific year. Due to the multi-annual nature of some programs, two categories of appropriation are entered into the EU budget: appropriations for commitments, i.e. the expenditure committed by the EU in a given year with respect to operations that can be carried out over a longer period of time, and appropriations for payments, i.e. the expenditure effectively incurred by the EU in a given year in meeting the commitments of that and/or of previous years.

The *principle of equilibrium* provides that the EU budget cannot be in deficit or surplus.

Practical enforcement of the principle is through automatic adjustment of the GNI revenue source to expenditures.

The *principle of specification* states that no commitment can be entered in the EU budget without a definite scope and purpose. The only exception is the budgetary reserve.

Finally, the *principle of the unit of account* states that the EU budget is expressed in EUR.

Procedure for the Annual Budget Adoption and Implementation

The procedures are comparable in many respects to procedures at the national level. The EU budgetary procedure consists of five main phases. Preparation of the budget for the year *N* starts with the *proposal of the Commission* submitted by the end of April of year $N - 1$. In the second phase, *Council and Parliament discuss the proposal and adopt the budget* with the required majority by the end of December of year $N - 1$. The next stage consists of the *Commission's execution of the budget* throughout year *N*. *Technical control of the budgetary execution*, which is in the hands of the European Court of Auditors, represents the fourth stage of the EU budgetary procedure and is usually completed around November of year $N + 1$ for the budget of year *N*. The fifth phase – *political clearance* – is given by the Parliament usually in March of year $N + 2$ for year *N*.

Procedures for the MAFF Adoption

The boundaries for the annual budget are set by the MAFF on the expenditure side and by the ORD on the revenue side, with both of them required to be adopted unanimously in the Council. This prerequisite of unanimity is one of the main reasons why MAFF negotiations usually turn into one of the most complex negotiations among the EU Member States, even at the European Council level, where political clearance has to be achieved. On the other hand, the unanimity rule of the Council *de facto* transfers the decision-making power to this institution. The negotiations about the MAFF 2007–2013 clearly confirm this fact, as the Council agreement was significantly changed by the Commission's proposal, and the Parliament's role in the decision-making process was rather symbolic. With the Lisbon Treaty in place, the importance of the Parliament in the MAFF decision-making process has strengthened substantially.

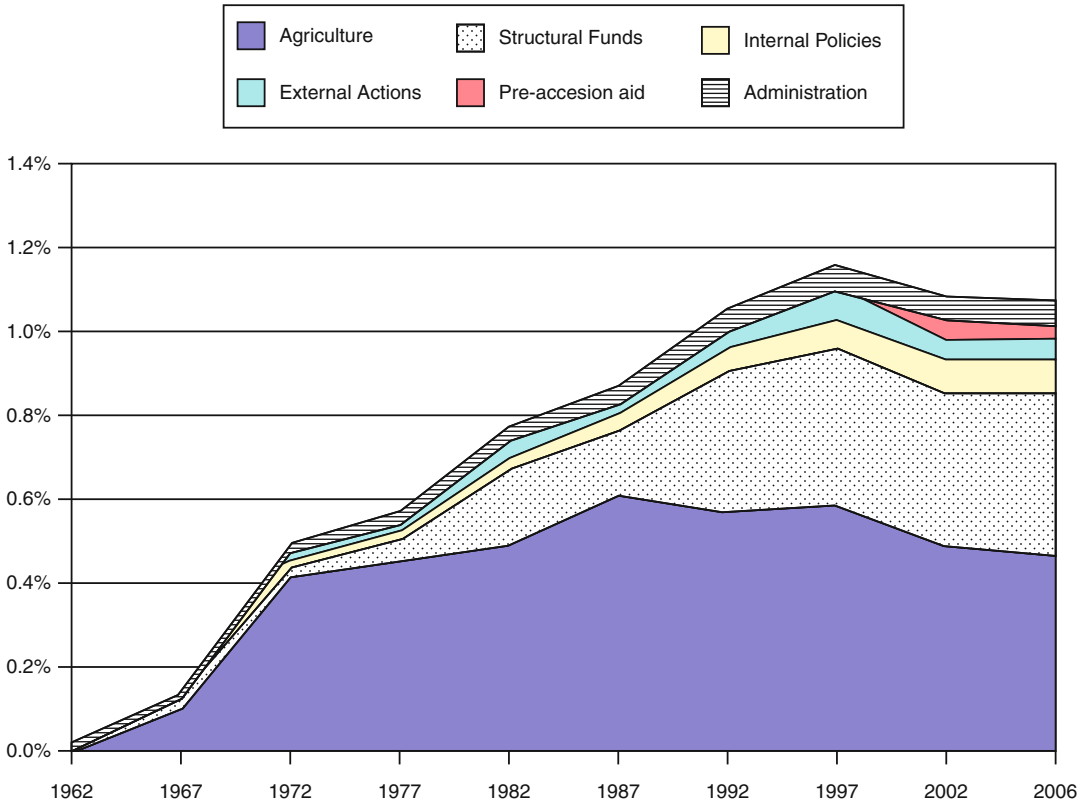
Structure of Expenditures

In the early decades of the Community, CAP absolutely dominated EU expenditure. More recently, due to several enlargements of the Community/EU and through the introduction of the MAFF instrument, CAP expenditures have been capped, allowing funding for some other items, especially cohesion policy expenditures. As shown in Fig. 1, approximately 80 per cent of all EU budget funds was earmarked for these two policies over the last two decades. The remaining share of the budget was allocated for external EU activities and internal policies aimed at boosting competitiveness and the implementation of other objectives.

In the MAFF 2007–2013, the total volume of finally agreed expenditures will amount to h864 bn in commitment appropriations and h821 bn in payment appropriations. As shown in Table 1, these amounts are significantly lower than the comparable figures proposed by the Commission.

The MAFF 2007–2013 classifies expenditures under six headings:

1. *Sustainable growth*, subdivided into competitiveness for growth and employment (research and innovation, education and training, trans-European networks, social policy, economic integration and accompanying policies) and cohesion for growth and employment (convergence of the least developed EU countries and regions, EU strategy for sustainable development outside the least prosperous regions, inter-regional cooperation);
2. *Preservation and management of natural resources* (common agricultural policy,



European Union Budget, Fig. 1 Evolution of EU budget expenditures 1962–2006 (as a percentage of EU GNP/GNI) (*Source*: European Commission)

European Union Budget, Table 1 EU budget expenditures under the MFAA 2007–2013 (*Source*: European Commission and author's own calculations)

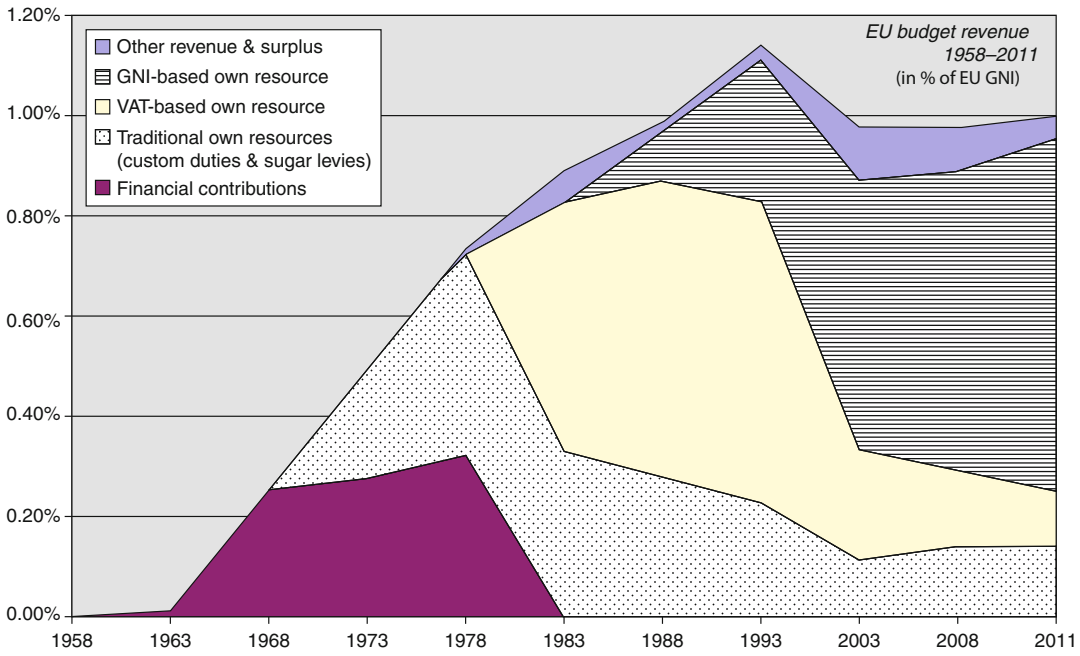
Expenditure headings (in commitment appropriations)	Commission proposal (February 2004)		Inter-institutional agreement (June 2006)		Change between June 2006 and February 2004
	h bn	%	h bn	%	Change (%)
1 Sustainable growth	462	45	382	44	−17
1 A Competitiveness	*122*	*12*	*74*	*9*	*−39*
1 B Cohesion	*340*	*33*	*308*	*36*	*−9*
2 Natural resources (CAP)	400	39	371	43	−7
First pillar	*301*	*29*	*293*	*34*	*−3*
Second pillar	*99*	*10*	*78*	*9*	*−21*
3 Citizenship, freedom, security and justice	21	2	11	1	−49
4 EU as a global player	85	8	49	6	−42
5 Administration	58	6	50	6	−14
6 Compensations	0	0	0,8	0	n.p.
Total	1.025	100	864	100	−16

common fisheries policy, rural development and environmental measures);
3. *Citizenship, freedom, security and justice*, subdivided into freedom, security and justice (justice and home affairs, border protection, immigration and asylum policy) and citizenship (public health, consumer protection, culture, youth, information and dialogue with citizens);
4. *EU as global player* (covers all external actions by the EU);
5. *Administration* (covers the administrative expenditure of all the European institutions); and
6. *Compensations* (includes compensatory payments relating to the latest expansion of the EU).

Structure of Revenues

The evolution of the revenue side of the EU budget has been driven by the continuous attempt to strike a compromise between the financial autonomy of the EU budget and sufficiency of resources for its financing. The budget of the ECSC in the early 1950s was financed through a tax-based 'own' resource (a levy on steel production) while, in contrast, the Rome Treaty stipulated that the EEC budget was financed in a totally intergovernmental way, i.e. through direct contributions from Member States. At that time, the EEC budget had no 'own resources' and thus had no financial autonomy from its Member States.

It was in 1970, with the Luxembourg Treaty, that the EEC budget started to move towards an own resources model. At that time, own resources included traditional own resources, i.e. customs duties and agricultural levies, and VAT-based revenues from Member States. This structure remained unchanged until 1984, when the Fontainebleau Council introduced the UK correction, which in fact means a reduction of the UK contribution to the Community budget. The last major change to the EU budget revenue side occurred in 1988 with the Delors I package, which introduced the fourth resource, known today as the GNP/GNI resource. Figure 2 provides a historical overview



European Union Budget, Fig. 2 Evolution of EU budget expenditures, 1958–2011 (as percentage of EU GNI) (*Source*: European Commission 2011)

of how the EU budget has been financed since 1958.

Since its last major reform in 1988 and thus also in the MAFF 2007–2013, EU budget revenue has been made up of four major resources:

1. *Traditional own resources (TOR)*, which include customs duties and agriculture levies. In the 2011 EU budget, this source contributes around 14 per cent of total EU budget needs and is on a downward path.
2. *Funds levied on the basis of value-added tax (VAT-based resource)*, defined on the basis of a statistically adjusted VAT base of the Member States. Its share of the EU budget is also small and accounts for around 11 per cent in 2011.
3. Funds levied on the basis of the Gross National Income of the Member States (the GNI-based resource), i.e. funds earmarked for balancing the budget measured in proportion to the GNI of every Member State. In 2011, this funding source participates with around 70 per cent in total funding needs of the EU budget.
4. The UK correction, as formally the fourth EU budget own resource but in substance terms a zero sum mechanism. It amounts to an annual level of around h3.8 bn in the years 2010 and 2011.

In addition to the four own resources described above, there are some other revenues that may also finance the EU budget. It should be underlined that they are small in size and of a non-foreseeable character. In the 2011 budget, the major source of these other revenues is a budgetary surplus from the previous year. An existence of other revenues reduces the volume of GNI contributions to be provided by Member States.

Correction Mechanisms and Net Balances Issue

An integral part of the own resources system is formally also the 'UK correction' as well as a set of 'corrections on this correction'. Introduction of this instrument dates back into early 1970s when the UK joined the EEC. At that time the UK was among the poorest Member States, but due to the EU budget expenditure bias toward the CAP the country had a negative net financial balance towards the EU budget. As this was considered unfair by the UK authorities, after intense negotiations with other Member States the issue was resolved at the 1984 Fontainebleau European Council through the so-called 'UK correction' arrangement, whereby the UK became entitled to a refund financed by all other Member States. The economic logic of the arrangement was that the UK position *vis-à-vis* the EU budget was excessively negative in relation to its level of development, and that the country is eligible for a rebate on its contribution to the EU budget.

Even though the UK position in terms of its economic development has improved substantially since 1984, eliminating (or at least substantially reducing) the justification for the 'UK correction', the system remains in place with only minor changes. This can be explained by the fact that the 'UK correction' is an integral part of the ORD, which requires unanimity for changes. The correction mechanism system in place in the MAFF 2007–2013 contains, in addition to the 'UK correction' corrections to four other large net payers to the EU budget, whereby Germany, Austria, the Netherlands and Sweden pay only 25 per cent of their normal 'UK correction' funding share.

There are also three other 'corrections on corrections'. First, Germany, the Netherlands, Sweden and Austria have an arrangement whereby their share in financing the 'UK correction' has been reduced via a reduction in the call rates for the VAT-based own resource.

Second, the Netherlands and Sweden receive a fixed lump sum reduction of their annual GNI contributions.

Third, Member States retain a fixed percentage of all traditional own resources collected. Since 2000 this percentage has been set at 25 per cent.

The Main Deficiencies of the EU Budget

The EU budget is dominated by a small number of highly redistributive policies, with CAP providing

funds for European farmers and cohesion policy redistributing funds towards less wealthy regions. Poor representation of broader EU-wide policies on the expenditure side of the EU budget associated with the domination of national contributions on the revenue side have resulted in a system particularly prone to bargaining and supporting the obsession of Member States to demand *juste retour* for their contributions to the EU budget. The two most recent MAFF negotiations have confirmed this pork barrel mentality, where achieving an acceptable net balance position has *de facto* become a more important objective than agreeing the size and structure of the spending.

Despite significant changes within and around the EU over the last two decades, the structural characteristics of the EU budget have remained largely unchanged since 1988. In addition to its strong pro-status quo bias, the EU budget has become less and less transparent. The 2007–2013 MAFF negotiations are a clear confirmation of these problems. In order to achieve acceptable net balances Member States were ready to sacrifice Lisbon-type expenditures as one of the top EU policy priorities at that time, and they also insisted on the continuation of the current corrections and rebates as well as the introduction of new ones.

Structural rigidities in the EU budget reflect the current institutional set-up of the EU and its decision-making system. The MAFF adoption procedure combines elements of an intergovernmental and a supranational approach. According to the Lisbon Treaty, an agreement on a proposal prepared by the European Commission has to be reached in the European Council by consensus, and in the European Parliament by majority. The two institutions have different incentives in these negotiations. While the European Council aims to reduce EU budget expenditures in order to reduce the Member States' contributions from their national budgets, the European Parliament has an incentive to increase the expenditures as the required funds will be provided automatically by the Member States up to the ceiling determined by the ORD.

There have been several attempts to address the net balances problem and the highly complex and non-transparent system of EU budget corrections, but to date none have been successful. The most serious attempt was the 2004 Commission's proposal for a generalized correction mechanism that would be open to all Member States and would replace the 'UK correction'. There have been other ideas to address this subject: One was to modify the calculation of net balances so as to take into account a broader concept of costs and benefits apart from pure budget. There are proposals from academic circles (see, for example, De la Fuente and Domenech 2001; Heinemann 2007; Rant and Mrak 2010) proposing that EU budget negotiations would be divided into two stages. In the first one, net budgetary positions should be fixed followed by the negotiations on the content of the EU budget in the second stage.

Before the 2014–2020 MAFF Negotiations

EU budget has remained conceptually unchanged since the 1988 reform in spite of the fact that the EU has undergone significant changes, including its enlargement from 12 to 27 Member States, completion of the internal market and introduction of new priorities in the areas of internal and external policies. Under the dominant influence of the *juste retour* logic, the main victim of the 2007–2013 MAFF negotiations was the Lisbon strategy itself, even though all Member States had explicitly supported international competitiveness as the top substantive priority of the EU in the forthcoming period. Being aware that the MAFF deal struck under these negotiations was not only unsuitable in its substance but also highly non-transparent in its financial terms, the December 2005 European Council authorized the Commission to prepare a thorough review of the EU budget and report back in 2008 or 2009.

Due to numerous political considerations, including complications with the Lisbon Treaty ratification process and the 2009 European Parliament elections followed by the appointment of the new Commission, the review was published only in late 2010. This delay has in fact prevented the review from serving its original purpose, i.e. to

provide a basis for a thorough discussion about the possible reform of the EU budget. In fact, the review, which was published less than a year before the onset of the negotiations about the 2014–2020 MAFF, has turned into nothing more than a good issue paper for the forthcoming negotiations. Unfortunately, a prime opportunity for a thorough reform of the EU budget has been lost and the negotiations seem to be burdened again with the *juste retour* logic and the dominance of net national budgetary positions.

During the 2014–2020 MAFF negotiations, the Member States and the EU institutions will have to address numerous challenges, including the relatively weak international competitiveness of the European economy, negative consequences of the ongoing financial and economic crisis, incompleteness of the internal market, continued social and economic disparities, import energy dependence and climate change. On the institutional side, this will be the first MAFF to be negotiated under the Lisbon Treaty, with a formally substantially stronger role of the Parliament in the process.

The official proposal of the Commission for the 2014–2020 MAFF negotiations issued in June 2011 has the following main characteristics:

1. The size of the EU budget is to be kept at a similar level as before, i.e. at a level of around 1.05 per cent of EU GNI in payment appropriations.
2. CAP and cohesion policy remain the two largest spending items in the EU budget, although the former will have a significantly reduced share in total expenditure compared with the 2007–2013 MAFF period.
3. On the revenue side, VAT-based resource is proposed to be abandoned while two new EU taxes – a financial transactions tax and a VAT tax – are proposed to be introduced as of 2018.
4. The existing 'UK correction' and 'corrections on corrections' mechanism is proposed to be simplified and exchanged by lump sum gross reduction of GNI payments for four large net payers into the EU budget, namely the UK, Germany, the Netherlands and Sweden.

See Also

- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ Euro Zone Crisis 2010

Bibliography

Altomonte, C., and M. Nava. 2005. *Economics and politics of an enlarged Europe*. Cheltenham: Edward Elgar.

Begg, I. 2005. *Funding the European Union*. Federal Trust Report on the European Union's Budget, London.

Begg, I. 2007. The 2008/2009 review of the EU budget: Real or cosmetic? CESifo Forum 1/2007.

De la Fuente, A., and R. Domenech. 2001. The redistributive effects of the EU budget: An analysis and proposal for reform. *Journal of Common Market Studies* 39(2): 307–330.

Enderlein, H., J. Lindner, O. Calvo-Gonzales, and R. Ritter. 2005. *The EU budget: How much scope for institutional reform*. Frankfurt: ECB.

European Commission. 2008. *EU public finances*. Brussels: European Commission.

European Commission. 2011a. *A budget for Europe 2020*. Brussels, 29 June, COM (2011)500 final.

European Commission. 2011b. *Financing the EU budget*. Brussels, 29 June, SEC (2011)876 final.

Heinemann, F. 2007. Solving the common pool problem in the EU fiscal constitution. Presentation at the conference Challenges to the EU Budgetary Reform, Ljubljana, May 7.

Le Cacheux, J. 2005. European budget: The poisonous budget rebate debate. In *Etudes et Recherches*, vol. 41. Paris: Notre Europe.

Rant, V., and M. Mrak. 2010. The 2007–13 financial perspective: Domination of national interests. *Journal of Common Market Studies* 48(2): 347–372.

European Union Single Market: Design and Development

Jacques Pelkmans

Abstract

What precisely is a Single Market, how it has been designed in the case of the European Union (e.g. in the treaty) and how it has

developed over 5 decades, are the three questions answered in this contribution. It is first shown that the design of a Single Market matters: it is not just about goods markets (despite the enormous emphasis in the literature on this aspect) but also about services, labour, capital and codified technology. In order to have a Single Market function properly, it is indispensable to combine negative integration (removal of barriers) with a considerable ambition in positive integration (common regulation, selected common policies, common market institutions where appropriate and endowed with proportionate but sometimes overriding powers). The treaty contained a unique design which has been 'upgraded' with the increasing ambitions of 'deepening' and 'widening' of scope of markets and policies in the EU. The development of the EU Single Market is stylized in four accomplished stages after the mid-1980s, when the 'customs-union-plus' was overcome for a much 'deeper' internal market, until today.

Keywords

European Union; Single market; Internal market; Economic integration

JEL Classifications

F15

The Roots of the Single Market

The European Union, which began as the European Economic Community in 1958, was based on the new idea of a 'common market'. Strictly, the Rome treaty did not define what the 'common market' was. However, most analysts at the time saw it as the combination of five 'economic freedoms' (four instances of 'free movement' (namely, for goods, services, capital and 'persons', probably including workers) and the right (of companies or individuals) to establish in any other EEC country, various forms of common policy-making (trade, competition, agriculture and transport) indispensable for a common market, common regulation, officially called 'harmonisation', and some coordination or lighter cooperation. All these aspects were specified to some degree in the Rome treaty. Later, with the first revision of the treaty in 1985, the notion of the 'internal market' was introduced in the text. The proposed treaty revision coincided with the famous EC1992 programme of 7½ years aiming for the 'completion of the internal market' (from mid-1985 to late 1992). The term 'completion' as well as the sheer ambition of the EC1992 programme quickly led many people to speak colloquially about the EU 'Single Market' ever since. The treaty has been revised four times since 1985, yet the term 'Single Market' is nowhere to be found. The present contribution will focus on the concept and treaty design of the Single EU Market, and will subsequently show how the EU internal market developed over time.

What Is a Single Market?

More Than the Law of One Price

The benchmark of a single market, which is suggested to every student of economics, is the law of one price. This textbook idea is a useful and simple summary indicator of the result that market integration will yield. Taking it literally would be misleading: even a local market in a small village does not exhibit complete price equality and whether price differentiation is a function of product differentiation on that local market is not easy to verify for consumers, and certainly not for incidental visitors without repeat purchases. Nevertheless, the benchmark is useful because it is expected that, in a single market, price divergences are held in check by actual and potential competition in that market and spatial competition from nearby markets in other locations. The assumptions behind such price convergence include good and timely information and actual and potential mobility of suppliers as well as consumers. Nevertheless, it reflects a very narrow perspective of market integration and its utility is more questionable in markets other than goods, such as all areas of services, in goods and services of network industries, in labour markets across

countries and in knowledge markets (regulated or not by intellectual property rights). Nevertheless, there is much more to a Single Market than a tendency towards price convergence. A Single Market is also about (quality and other) differentiation in goods, services, capital and labour – hence, the gains of variety – and about the stimulus to come up with innovative ways of engaging in competition, be it via goods or services or process innovation, or in distribution or marketing. Companies long used to being a major player might be challenged by different 'business models' more appealing to the same or new consumers or customers. In some sectors, initial national prices may be a bad predictor of later prices in the Single Market due to scale economies (perhaps even amplified by 'learning curves') which can only be reaped with much larger volumes of turnover (see Pelkmans (2011) for a brief survey of the economic impact of the EU Single Market). For all these reasons, price convergence is a helpful but insufficient indicator of market integration for economists. For a proper economic understanding of a single market, one should appreciate the driving forces behind the eventual economic gains of a Single Market: competition in static and dynamic forms driving not only price convergence but also cost minimization and greater variety in goods and services, innovation and choice.

There is also the crucial issue of single market design. Does price convergence and a rich view of competition answer the query 'what' an internal market is? In fact, it does not; it merely tells economists whether a prominent economic test of the 'working' of that market is satisfied. If European integration has taught one lesson, it is that the building of such a Single Market is an extremely complex, highly intrusive and staggered undertaking. And this matters a lot for the economic study of the Single Market. The large 'distance' between one basic economic criterion to assess a Single Market and the many stages of its complex development has inevitably generated a literature which mainly focuses on general and 'aggregated' outcomes. Typically, it fails to give much economic guidance on the what and how of (deep) market integration.

Therefore, when going from the general concept of the internal market to a practical design which can be used for a treaty, the question of what a Single Market is usually answered by a 'stages theory of economic integration'. What does it take, in terms of measures of the EU and market institutions, to get a well-performing Single Market? The traditional institutional approach was initiated by Balassa (1961) and was much refined and adapted later in the light of the EU experience (see Pelkmans 1982, 1985; see also Lloyd 2005). The five Balassa stages are: free trade area, customs union, common market, economic union and total economic integration. Whereas the first two were taken from GATT, art. 24, the other three are new concepts. With the latter three, there are serious problems of design logic.

Single Market: Stages Beyond a Customs Union

The 'common market' – beyond the free circulation of goods in the customs union with common tariffs – is defined by Balassa as the free movement of factors of production (capital, labour, and nowadays also codified knowledge, as in patents etc.). His common market has no institutional features other than a legal duty to liberalise cross-border flows of these factors. Clearly, this is a fantasy world: cross-border free movement of factors would at the very least assume common regulation for labour (not to speak of the profound implications for the welfare states) and for codified technology or knowledge but almost certainly it would also require common institutions enabling more detailed decision-making (especially for harmonisation and EU regulation). In the early 1960s most European countries still had fixed exchange rates and exchange controls; hence cross-border freedoms in capital movements would have had major macro-economic implications. In other words, 'negative' market integration (only removal of barriers) in a common market is unthinkable without ambitious 'positive integration' (common regulation, or lighter harmonisation, plus coordination and

some common decision-making in common institutions) (see also Tinbergen 1954). If countries are wary of this degree of centralisation or commonness, positive integration will be insufficient, and without it 'negative' integration in factor markets will become impossible: it will simply not happen. Another gap in Balassa's third stage is the neglect of services. Again, since many services are regulated and several are tightly supervised, a single market for services cannot come about without common regulation (overcoming market failures, ideally) and some central or coordinating supervisory agencies. As services cover a huge set of economic activities, the scope of centralization would increase considerably. But even in goods, a common market is far more than just free circulation behind common tariff walls. Nowadays, it is well understood that there is a broad range of other barriers or trade cost-raising elements in goods markets which have to be addressed if there is to be any chance of obtaining a 'single' market. But this simple fact changes radically the nature and ambition of that single market. The EU led the way in addressing such barriers, after having dealt with the customs union aspects without any remaining exceptions in less than the required 12 years. But the various 'non-tariff barriers' appeared to be much more difficult to tackle and eventually prompted innovative ways to overcome the frustrating stalemates in Council. At the same time, this process also led to degrees of regulation (and co-regulation, for example in overcoming technical barriers) and selective centralisation which had not been expected by the founding fathers.

The fourth stage of Balassa is called 'economic union'. In Pelkmans (1991), a literature survey since the late 1940s shows that economists have not been disciplined in utilizing a single definition: no fewer than seven definitions, with very different meanings, can be found. Balassa's definition brings in the common institutions and 'positive integration' only at this stage (rather than already for the common market), yet remains vague about what exactly the economic purpose and scope of this 'union' should be. If indeed the disparities in national policies would lead to discrimination (of market players or their goods, services etc.), it would have to be dealt with at the common market stage. As we shall see later, this kind of 'economic union' cannot serve an eventual monetary union either. This simple point underscores that the treaty design of deeper economic integration ought not to be taken lightly by economists.

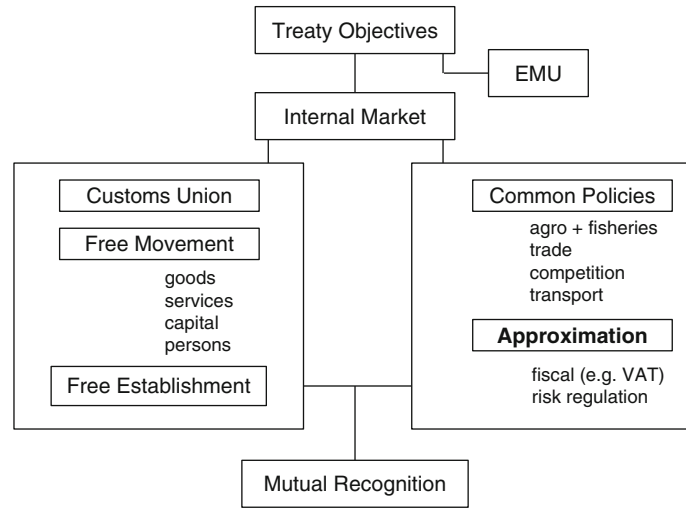
The fifth Balassa stage is called 'total economic integration'. It has two key elements: the 'unification of monetary, fiscal, social and countercyclical policies' and the 'setting up of a supranational authority where decisions are binding for the Member States'. The latter element is several stages 'too late' since substantial positive integration will be required already for genuinely free movement of goods (beyond free circulation in the customs union), and the more so for services, labour, capital and codified technology in the common market. The former element (unification) is not only a huge jump from a vague 'economic union' but there is no obvious justification for so much centralisation. A 'subsidiarity test' (Pelkmans 2005) boils down to a cost–benefit analysis of (de)centralisation of public economic functions such as monetary, fiscal and social policies given an already realised single market. With centralisation criteria like scale and cross-border externalities, besides decentralisation criteria such as diversity of preferences among regions or countries and the (*ceteris paribus*) greater ability of local politicians to 'read' such preferences and act accordingly at the local level, a subsidiarity test will yield a much more nuanced view than Balassa's fifth stage. It would show that social policy is highly unlikely to be a candidate for centralisation, and fiscal policy is suitable only in some respects (e.g. debt caps, but no or only modest union taxes). For present purposes, one can also ask the fundamental design question of whether a 'deep' single market can perform well over time without monetary union or without at least a credible mechanism for maintaining stable exchange rates.

Altogether, a Single Market is a highly ambitious 'means' for the pursuit of higher economic (and possibly non-economic) aims.

The EU's Single Market Design in the Treaty

Today, the relevant treaty for the EU Single Market is the Treaty on the Functioning of the EU (TFEU), in force since 2010. It is one of the two Lisbon treaties (the other is the EU treaty). The basic idea of the internal or common market (the treaties have never used the term 'Single Market') has not changed much over time. What has happened is that both decision-making and 'deepening' have been facilitated several times. This has turned out to be critical. Figure 1 stylizes the EU internal market idea found in TFEU and case-law of the EU Court (CJEU). Figure 1 expresses the notion that the internal market is a 'means' – indeed, the principal means – for the pursuit of EU treaty objectives. Ever since the Rome treaty, these objectives prominently feature economic growth (in various formulations and under conditions, e.g. sustainability). Recurrent treaty revisions have overloaded the EU with new objectives, but how crucial these are and what priority they have is unclear. In any event, every time the Single Market moves (back) to the top of the EU's agenda, the main motivation was and remains economic growth and/or productivity hikes. Figure 1 has a symmetric set-up: on the left side one finds 'negative market integration' (customs union, the four free movements and the right of establishment) and on the right hand side 'positive integration' consisting of the most important common policies as well as two prominent forms of 'approximation' or harmonisation (risk regulation, comprising a very large part of EU regulation and related to health, safety, environment, consumer and saver/investor protection; indirect tax regulation on e.g. VAT). In addition, Fig. 1 depicts 'mutual recognition' hanging in between the two.

Of course, with initial veto-based decision-making and the greatest hesitation on the part of the EU Member States to radically pursue cross-border intra-EU liberalization in goods, services, labour, capital and codified technology, as well as to apply free establishment to all sectors, or to engage in far-reaching and often intrusive risk regulation in many areas and submarkets as a condition of free movement, one can understand that the institutional state of EU market integration has only gradually moved towards the ideal picture of Fig. 1 over a period of five decades. This also goes for the common policies: initially, when a common agricultural policy (as the basis for an internal market in agro-goods) was created, there was no equivalent in fisheries; the common transport policy in the six modes only came about in earnest during the second half of the 1980s; EU competition policy operated almost 30 years without merger control, which is now the most



European Union Single Market: Design and Development, Fig. 1 EU internal market in the treaties

important aspect; and EU trade policy in goods was only complete once third-country quotas at the national level had been outlawed (in 1992), whilst in services and investment full EU-level power was only granted in the Lisbon treaty. All such difficulties are good illustrations of why a proper understanding of the EU Single Market and its actual or potential economic impact is so demanding. Indeed, while Fig. 1 serves as a guide for the overall concept, a proper understanding necessitates the zooming in on some of the more important details and how these EU accomplishments have accumulated over time.

Deepening, Widening and Enlargement of the EU Single Market Over Time

Given the incredibly broad range of negative and positive integration the Single Market entails, we shall stylise the progressive accomplishments of the Single Market in four steps: one on the state of achievements after 25 years of EEC; one on the EC1992 programme; one on the accomplishments between 1993 and 2010; and one on the 2011 Single Market Act. It provides powerful and concrete evidence of the rising ambition of the EU's Single Market. The increasing importance of the Single Market is the result of 'deepening' (firmer application of existing commitments, with fewer exceptions), 'widening' (of scope, that is, more domains are brought under the internal market) and 'enlargement' (more EU countries, hence a larger market size).

The EEC Common Market After 25 Years

Table 1 reflects what the EEC's 'common market' looked like in 1982. It is best described as a 'customs-union-plus'. The main items in the first word column are taken from Fig. 1. The table largely speaks for itself. The drafters of the EEC treaty and subsequently the Member States (and to a lesser extent, even the Commission) simply had no well-informed idea of what a common market, as specified in the Rome treaty, really requires.

Deepening and Widening Under EC-1992

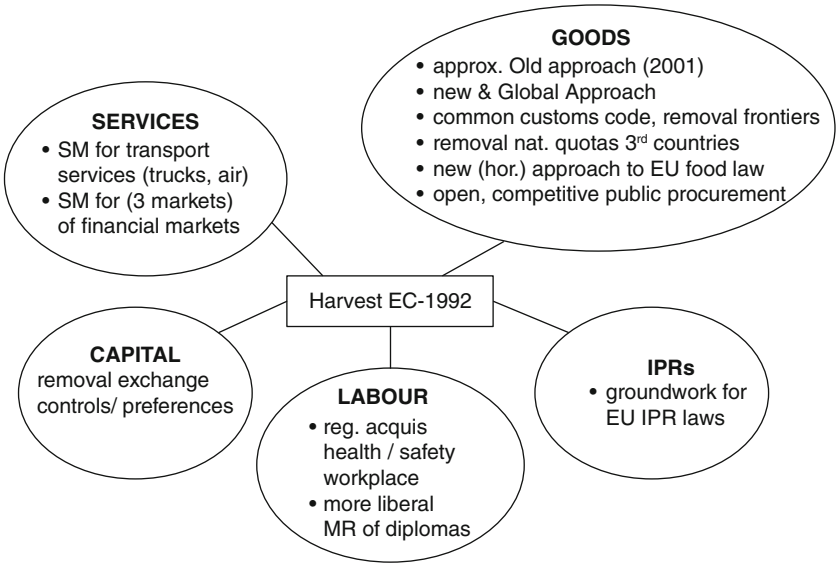
Table 1 allows one to appreciate how radical EC1992 was. The striking difference between Table 1 and Fig. 2 (see above) is that measures and new accomplishments are observed in all five areas of free movement. Besides a range of initiatives in goods markets, two big services sectors (transport and financial services) are tackled and exchange restrictions are abolished. In addition, some horizontal aspects were improved, such as merger control, the gradual inclusion of network industries and more emphasis on mutual recognition. The list for goods markets is impressive: the very detailed customs code is realised, inner frontiers and physical customs are abolished, a huge number of highly technical directives and regulations in risk regulation are adopted (Old and New Approach etc.) including conformity assessment (Global Approach); in addition, food law is remodelled based on mutual recognition, extensive SPS-type harmonisation in no less than 160 directives and the avoidance of food-specific directives (but horizontal food directives instead, with far lower costs while maintaining the great variety that national traditions cherish); finally, public procurement was addressed so as to become competitive and open EU-wide. This is no longer a customs-union-plus: it implies far-reaching free movement in goods and selectively in some important services, such as several modes of transport as well as free movement in and EU regulation of financial services. It also comprises free movement in (financial) capital (albeit that equity in stock exchanges is still subject to problems of clearing and settlement) and at least some minimum rules (e.g. in occupational health and safety) in labour markets. It does not yet add up to a common market, since services are only selectively addressed, network industries still have to be brought in and IPRs have not yet been resolved, for example. Moreover, the regulation and cross-border liberalisation of financial services (banking, insurance and investment services) was incomplete, with serious gaps and omissions due to the reticence of Member States,

European Union Single Market: Design and Development, Table 1 EU common market after 25 years

		Acquis	Gaps/omissions
1	Customs union	• CET	• No common customs code
		• No intra-EU quotas	
2	Free movement		
	2.1 Goods	• Case law on mutual recognition	• Very limited progress on technical barriers
			• No EU regime for national public procurement
			• Network industries excluded from common market
			• Very limited progress on regulatory barriers
			• Customs controls remain
	2.2 Services	• Minimal recognition of services free movement	• Almost no EU regulation in services
	2.3 Workers	• 1968 free movement decision; only 'residual' mobility	• Host of barriers to immigration (health, insurance, housing, taxes, pensions)
			• Many difficulties for frontier workers
			• Host country control removes incentives for migration
	2.4 Capital		• Six out of 10 EU countries maintain exchange controls
			• 1982 court case, allowing restrictions on capital flows
3	Right of establishment	• 1962 liberalisation of FDI	• (restrictive) licensing and authorisation in some (services) sectors
		• National treatment	
4	Common policies		
	4.1 CAP	• CAP well-established but highly distortive	• No fisheries policy
			• 'Green' exchange rates
			• No risk regulation (and controls) in SPS-type aspects of agriculture
	4.2 Trade	• Common policy, for goods only	• Some EU countries maintain selective quotas *vis à vis* specific third countries (e.g. Japan)
			• Nothing on services, IPRs, investment
	4.3 Competition	• (Narrow) anti-trust in place	• No merger control
		• For goods only	• *De facto,* no cases in services
			• Weak and inconsistent control of state aids
			• *De facto,* no cases in network industries
	4.4 Transport	• Some harmonisation (e.g. technical)	• Trucking subject to selected quotas
			• No cross-border liberalisation in air, rail, buses, maritime
			• Rigid restrictive regime in river transport
5	Approximation		
	5.1 Fiscal	• VAT base, regime and bands of VAT rates	• 'Trade' costs at frontiers still high
		• Excise duties on only three types (others illegal)	
		• Case law on alcohol taxes	
	5.2 Risk regulation	• Some selected results in regulation of high-risk goods (Old Approach)	• Vast areas of risk regulation of goods still national and disparate
			• No European standards, except (some) electrical
			• No EU rules for conformity assessment
			• No risk regulation for services or labour/hence no EU supervision

NO EU IPRs or harmonisation

Note: 'Acquis' is a term used in the EU for the accomplishments at EU level



NOTES: (1) CJEU case law promoted M.R.; (2) preventing new barriers in SM via 83/189; (3) merger control '89; (4) network industries not in White Paper (exc. broadcasting); began early 1990s.

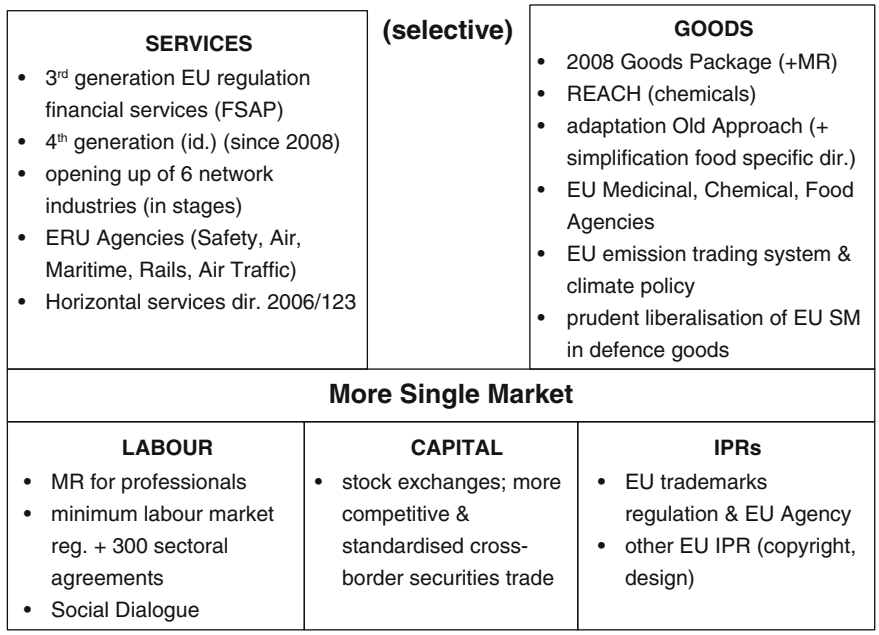
European Union Single Market: Design and Development, Fig. 2 What EC-1992 accomplished

whilst supervision was kept at national level, based on fairly general EU regulatory principles.

Incremental Deepening After 1992

Some further progress as well as refinement of the internal market acquis has been accomplished since 1992. It is summarized in Fig. 3.

The main elements of deepening consist in IPRs at EU level (except the EU patent) and a major assault on barriers in services markets, in a number of ways. The horizontal services directive marks a U-turn in that all services markets should benefit from free movement, unless already regulated by EU rules (or exempted in a few instances). The other significant progress is liberalisation in six network industries (telecoms, electricity, gas, rail, air and postal). Also, a third generation of financial services regulation was built up, after the rules for establishment of financial institutions in the 1970s and those for free movement and mutual recognition of national supervision during the EC1992 process. The third generation (2000–2006) made EU financial services regulation more complete and technically more refined, in particular for investment services. However, it was insufficiently realised that the quality of EU regulation suffered from the undue emphasis of 'light touch' by the London City and the eagerness of many banks and others to exploit financial innovation. The EU regime did not guarantee that market failures were fully overcome and delivered insufficient guarantees for the proper assessment, management and pricing of risks in financial markets, especially at the wholesale level. In the event of major mistakes by large financial players, supervised or not, bank failures would become a real possibility, in extreme cases leading to contagion and systemic risks, hence endangering EU financial stability. Contagion was no longer a theoretical possibility, since the interconnectedness of banks in the deepened EU financial markets was increasing very rapidly. Furthermore, although national supervisors became embedded in EU supervisory networks, the latter remained cooperative and had neither the authority to act directly at EU level nor precautionary plans and/or funds to address cross-border contagion and systemic risks. Actual market



NOTES: (1) modernisation of EU competition policy; (2) RIAs (since 2003) and Better Regulation; (3) better inter-MS horizontal / adm.cooperation; (4) public procurement, 2nd generation.

European Union Single Market: Design and Development, Fig. 3 Deepening and widening the EU internal market: 1993–2010

integration began to run much too far ahead of the EU regulatory regime.

Once the financial crisis broke out in the autumn of 2008, the EU changed course: a fourth generation of EU financial regulation has been built up under duress (2008–2011). One may characterise this fourth generation by three key words: 'better quality' of regulation in overcoming market failures (e.g. in banking) by means of higher capital requirements or otherwise and doing away with 'light touch'; regulating 'all' financial activities or actors, thereby including credit rating agencies, hedge and other investment funds and derivatives; and shifting to 'more centralization' in supervision, via EU Agencies, and in closely monitoring systemic risks and financial stability in a special EU Board. In addition, bank resolution rules have been tightened, with explicit shareholder risks and minimising risks for deposit holders, in combination with proposed EU funds which can immediately address cross-border bank failures if necessary.

In goods markets, the only widening of scope is in defence goods, where restrictions led to absurd practices and trade costs. All other initiatives amount to refinements, either by protecting better free movement and EU rules (e.g. the 2008 goods package), or by improving the benefit/cost ratio of existing EU regulation (e.g. REACH; a more flexible Old Approach) or by joint technical expertise in EU Agencies. Regulation related to climate strategies is new and rightly allocated (given cross-border externalities) at EU level. Mutual recognition of diplomas – a difficult issue even in federations – has been attended to, but this thorny question will require a much more thorough EU approach before it will effectively alter conduct in markets. Moreover, the benefit/cost ratio has also improved for EU competition policy (via modernisation) and for the quality of (proposed) EU regulation (via regulatory impact assessment = RIAs). Going by the list in Table 1, the EU has meanwhile shifted a lot closer to an advanced form of an internal market: how 'single'

it is or not is crucial for the appreciation of further deepening and widening. One obvious gap is the lack of a European labour market, something a federation will enjoy but which would be possible in the EU, if and only if potential mobilities across intra-EU borders could be much larger. Only then would one be able to make an economic case for assigning labour market regulation and (at least part of) social transfers to the EU level. However, that unlikely outcome would, in turn, necessitate EU social charges or taxes, for which the treaty provides no legal basis. It is also possible to identify many 'hidden' obstacles, for a truly 'single' market to emerge in areas other than labour. This even includes a typical taboo so far in European integration, namely EU infrastructure. Certain network markets (electricity, gas, rail) cannot function properly on an EU scale without modern infrastructure designed with the single market in mind, not national or local priorities.

The Single Market Act

The fourth phase, initiated with the Single Market Act [COM (2011) 206 of 13 April 2011, Single Market Act] is yet another attempt to further deepen the internal market, in two steps, the first of which has been specified in some detail with 12 'levers'. The plan comprises initiatives to remove barriers – some of them deep-seated – as well as facilitation and cost-reducing measures which render the use of the Single Market more attractive. Given the high hurdles for smaller companies to actively participate in the internal market (see e.g. Mayer and Ottaviano 2007) these measures may well induce a higher degree of market integration. Amongst the measures removing obstacles, it is worth mentioning the EU common patent (and laws and a Court for common EU patent litigation) where the pure cost reduction of a patent would be as high as 80% and, perhaps even more importantly, the stimulus for innovation would be considerable (Guellec and van Pottelsberghe 2007) and permanent. Other measures proposed include greater investment in (mostly cross-border) energy and transport infrastructures, the removal of complicated obstacles to the Digital Single Market (often linked to private – e.g. contract – law and divergences in consumer protection, which typically tend to have remained largely national) and removal of (e.g. tax) barriers to a Europeanisation of venture capital. A full accomplishment of the Digital Single Market is expected to yield an increment to EU GDP of up to 4% (Copenhagen Economics 2010).

Conclusion

Altogether, the EU Single Market has steadily deepened and widened (in scope) over time. It has also become much larger, from 6 to 27 participating countries. Starting with a uniquely strong treaty, still one of a kind in the world, it has nevertheless taken some five decades, several critical treaty upgrades, and a host of programmes and special initiatives in order to arrive at a 'deep' but nonetheless incomplete Single Market. It has become an impressive edifice which, in some respects, is equivalent to internal markets in federal countries (Anderson 2011) and in other respects exhibits significant shortcomings. What is lacking in particular is a common labour market and the prospects for that are dim. Labour migration inside the EU will remain dominated by east–west flows (but only in a few professions, mostly low skilled) as long as wage divergences remain significant, and otherwise will continue to be 'residual'. The other weak element – a common market for services – has been tackled recently with greater drive and intensity, but it would not be surprising to expect another one or more programmes to arrive at deeper services market integration in the coming decade or so. Nowadays, many services markets are still fragmented in the EU despite the 2006 horizontal services directive, and their contestability is often weak as well. Enforcement by the Commission, new business models in some submarkets and consolidation, as well as new entry, will inevitably take time. Also, services which fall outside this horizontal approach, such as network services, professional services and retail finance (mortgages or consumer credit for example), and also employment agencies and freight rail, are

still encountering numerous problems when Europeanising their business strategies. Financial wholesale markets are rather well integrated. The remaining problems of the fourth generation include queries about (insufficient?) centralisation of supervision and crisis management, besides a few lingering doubts about EU regulation (e.g. should risks of banks arising from large exposure to large private parties not be extended to public securities as well?). A final weak spot in the Single Market is public procurement, a giant market in the EU (some 16% of GDP), where recorded cross-border contracts remain disappointingly low, despite several revisions of directives and reduced red tape.

See Also

- ▶ Debt Mutualisation in the Ongoing Eurozone Crisis – A Tale of the 'North' and the 'South'
- ▶ European Central Bank
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Cohesion Policy
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ European Union Budget
- ▶ European Union (EU) Trade Policy

Bibliography

Anderson, G. (ed.). 2011. *Internal markets and multi-level governance*. Don Mills: Oxford University Press, forthcoming.

Balassa, B. 1961. *The theory of economic integration*. Homewood: Irwin.

Copenhagen Economics. 2010. *The economic impact of a European digital single market*. Study for EPC. http://www.epc.eu/dsm/2/study_by_Copenhagen.pdf

Guellec, D., and B. van Pottelsberghe. 2007. *The economics of the European patent system*. Oxford: Oxford University Press.

Lloyd, P. 2005. What is a single market? An application to the case of ASEAN. *ASEAN Economic Bulletin* 22(3): 251–265.

Mayer, T., and G. Ottaviano. 2007. *The happy few – The internationalization of European firms*. Brussels/London: Bruegel and CEPR.

Pelkmans, J. 1982. The assignment of public functions in economic integration. *Journal of Common Market Studies* 21: 97–125.

Pelkmans, J. 1985. The institutional economics of European integration. In *Integration through law – Europe and the American federal experience*, vol. 1, Book 1, eds. M. Cappeletti, J. Weiler and M. Secombe. New York/Berlin: Walther de Gruyter.

Pelkmans, J. 1991. Towards economic union. In *Setting EC priorities, 1991–1992*, ed. P. Ludlow. London: Brasseys.

Pelkmans, J. 2005. Testing for subsidiarity. In *Die Europaeische Union: Innere Verfasstheit und globale Handlungsfaehigkeit,* ed. T. Bruha and C. Nowak. Baden Baden: Nomos.

Pelkmans, J. 2011. European union single market: Economic impact. In *The new Palgrave dictionary of economics*, forthcoming.

Tinbergen, J. 1954. *International economic integration*. Amsterdam: North-Holland.

European Union Single Market: Economic Impact

Jacques Pelkmans

Abstract

This compact literature survey covers the economic impact of the old EEC customs union and, more extensively, of the Single Market as it has emerged since the mid-1980s. The emphasis is on micro-economic studies of the effects on trade in goods (initially trade creation and diversion) and degrees of market integration measured in various ways, such as 'home bias'. Aspects discussed also include trends of price convergence, effects on competition, and induced impact on static and dynamic efficiency (e.g. innovation) as well as variety and higher productivity at the firm level. Economic research on the internal market in services has barely begun in earnest. A very brief discussion of methods highlights some of the problems. Finally, the principal work on macro-economic effects is summarised; that is, on overall productivity,

economic growth or one-off effects of specific internal market initiatives.

Keywords
Economic integration; European integration; Internal market; Single market

JEL Classifications
F15

Introduction

The EU has built up an ever more ambitious internal market, colloquially called the Single Market, over a period of almost 45 years. For a good understanding of economic impact studies, it is essential to know what the state of the internal market was in the relevant period. For a stylised overview distinguishing stages and some details, see Pelkmans (2011). The following micro-economic effects of the Single Market can be considered: the degree of market integration, including variety, price convergence trends, effects on competition, efficiency effects including dynamic efficiencies such as innovation and higher productivity at the firm or sector level. Of course, the EU has become a very open economy (except for some temporate-zone agro-goods) also in services, and hence the global dimension renders Single Market effects and their attribution (to the Single Market or other factors including globalisation nowadays) more difficult. The literature has also been interested in macro-economic effects such as economic growth or growth of productivity, or one-off effects of specific internal market initiatives on GNP. The final section draws the main conclusions.

Economic Impact of the EEC Customs Union

The initial EEC-6 period was dominated by the effects of the EEC customs union, via extremely simple simulation (e.g. Johnson 1958) or *ex post* empirical quantification (e.g. Verdoorn and Schwartz 1972; Prewo 1974). Scholars were only interested in trade flows and welfare gains, starting from basic partial equilibrium customs union theory. Given this theory, little or no attention was paid to competition, price convergence, technical efficiency improvements (despite Scitovsky's (1958) profound qualitative analysis of the deep inefficiencies in the EEC) or effects on FDI (despite Scaperlanda (1967)). The effects on trade flows attracted most attention, with trade creation and trade diversion being separated out by many authors. Surveys include Mayes (1978) and Winters (1987). The thrust of this impact literature was that trade creation in manufactures was estimated to be much larger than trade diversion, but not in agriculture, where diversion was considerable. The induced intra-EEC trade varied, but findings could go as a high as 50% (of 1958 trade) or more when controlled for growth and other factors. The methods employed relied on *ex ante* and *ex post* approaches, those with residual imputation (using an anti-monde of the same relations but without the EEC) and a range of distinct techniques. The welfare gains were computed as modest at best: the range was between 0.15% of GDP and 1% of GDP. Given that tariffs were still relatively high in the 1960s, such gains are disappointingly small. In part, this is due to overly simple models. And also to the fact that several, presumably significant, gains were omitted. Amongst these one may include (a) the reduction of X- (or technical) inefficiencies due to competitive pressures induced by the EEC; (b) scale economies (which require a sectoral approach); (c) gains from greater variety under intra-industry trade; and (d) terms-of-trade effects (which Petith (1977) estimates at 1% of GDP). Of these, technical efficiency improvements are likely to be the largest, but they risk being amalgamated in overall productivity studies with a range of determinants other than the EEC directly. In some sectors, scale economies led to spectacular welfare gains (e.g. washing machines and refrigerators; see Mueller 1981) as well as an upheaval of prior industrial structures.

Richer Economic Impact Studies of the Single Market

When the internal market – rather than the mere customs union – became a prominent issue, empirical studies and simulations moved away from the simple customs union framework. This was also true for the impact studies of some of the earlier EU enlargements (Viaene 1982, on Spain; Miller and Spencer 1977, a general equilibrium model on the UK; Gasiorek et al. 2002, also a general equilibrium approach, with simulation, on the UK accession; all of these focus on goods only) as well as for the Eastern enlargement (see Baldwin et al. 1997; Breuss 2002, using CGE simulations; and Breuss 2009 for a survey). Following the launch of EC1992, which was about 'deepening', many attempts were published to estimate the economic impact. Although EC1992 is complex and its follow-up even more so, the crucial additions consisted in major services sectors (all transport and financial markets) and, in goods, the removal of a huge number of technical or regulatory barriers and an opening up of public procurement. There have been two large-scale studies commissioned by the European Commission. The *ex ante* Cecchini report (Emerson et al. 1988; European Commission 1988) and the *ex post* Monti report (European Commission 1996, 1997. The Cecchini report made a breakthrough for three reasons: (1) for the first time, it attempted to use a (partial equilibrium) scale and imperfect competition approach and, in doing so, studying the 'pro-competitive' effects of the EC1992 approach, removing many regulatory and technical barriers; the Cecchini Group exploited ample funding for detailed and extensive fieldwork yielding data in areas which had never been addressed empirically (like public procurement in thousands of contracts) as well as for sectoral work, including selected services; and (3) it also attempted a macro impact simulation. With respect to the latter, when aggregating all the micro-effects into four categories and inserting them as 'shocks' (e.g. price reductions mainly) into two macro-econometric models of the European economy, the effects on GDP (a one-off increment), on employment after adjustment and on the price level could be generated. Based on the assumption of an exhaustive implementation of the 1985 programme, it found a 4.5% addition to GDP, an eventual 1.8 million extra jobs and a modest reduction in the price level.

We shall highlight two issues. First, what does exhaustive implementation mean? The Monti report finds lower empirical results than Cecchini and one of the reasons might well be the long-drawn-out and complex implementation issues. Indeed, the seven-year programme led to considerable implementation delays, partly due to slow action by Member States, but partly, too, because major areas required massive follow-up work after EC directives had been legislated. An example is found in the reliance on European standards. The New Approach (see Pelkmans 2011, Figure 2) sets essential safety (etc.) objectives in directives to overcome market failures, and refers to European standards written by independent European standardisation bodies but instructed (in EU mandates) to incorporate these objectives. Thus, the 1989 machinery directive is expected to require up to 1200 European standards and the completion of this huge and complex programme may well take 15 or more years. It goes without saying that, in the early 1990s, for which the data were used for the Monti report, only a few machine standards were in place, even though the one (!) item in the 1992 programme (a machines directive) had been properly dealt with. On this account, for Monti et al. to come to grips with the effects of these New Approach directives and their standards, the work should have taken place at least a decade later. Another reason for lower results by Monti is that the effectiveness of some initiatives (e.g. public procurement) left much to be desired.

Second, the Cecchini report deals with the pro-competitive effects of a deeper internal market in two stages: one is that the reduction of 'trade costs' increases import competition in the Single Market, reducing price-cost margins, and, given scale estimates, these will prompt larger output per firm, hence lowering average costs (although doubts about the empirical validity and relevance of scale were expressed by e.g. Geroski 1989); the other, more controversial, step in Emerson

et al. (1988) is that still 'deeper' integration is reflected by the assumption that prices converge; that is, full exposure to EU competition yielding output effects of up to 5% in concentrated industries. Subsequently, the latter effects were inserted as a shock into the model for the European economy. For many, this implies an upward bias in the report. On the other hand, the Cecchini report missed out on a range of 1992 issues which may have caused an opposing bias, including the abolition of exchange controls, the impact of creating EU IPRs (such as trademarks and copyright) and the removal of national VERs and quotas for some third party countries in, for example, cars and clothing. The most important underpinning of the Monti report is the study by Allen et al. (1998), which uses both CGE and econometric methods, but might nevertheless be somewhat problematic when focusing on the long term (e.g. Sorensen 1998). They also find selected sectors with significant extra output and a clear sharpening of intra-Single Market competition, measured by price–cost margins. The further assault on barriers preventing the EU Single Market in services from being realised began in earnest only with the 2006 horizontal services directive, and reliable empirical estimates are still awaited (but see De Bruijn et al. 2008).

The merits of newer empirical approaches such as CGE simulations and gravity have been called into doubt. Panagariya (2000) is particularly critical of CGE approaches, with and without the Armington assumption and given the utility functions employed. The *ex post* approaches increasingly rely on gravity equations with unclear and somewhat arbitrary links with economic integration theory, mainly because the alternatives are too demanding. Typically, in gravity approaches, detailed internal market effects are not specified but dummies for EU membership are used instead. One consequence is that gradual steps in the long-term economic integration process (such as the EC1992 programme) can no longer be empirically distinguished. A rare exception is Egger and Larch (2011), who arrive at empirical estimates of the economic impact of the Europe Agreements for Central European countries with the EU-15, before EU membership (the mid-1990s). They find that these Agreements generated some 5% extra GNP for the Central European countries and some 30% extra trade with the EU. They also show a strong redirection of intra-Central European trade, at first still influenced by defunct Comecon structures, towards bilateral exchange with the EU of no less than 50%! The authors rely on a structural model of bilateral trade inspired by Anderson and van Wincoop (2003) and utilising a Poisson pseudo-maximum likelihood estimation – thereby accommodating a large number of bilateral zero trade flows.

Conversely, general equilibrium analysts are critical of the popularity of trade creation and diversion, since these concepts are only meaningful in a carefully specified partial equilibrium context. This context is simply too restrictive as the elaborate survey by Panagariya (2000) shows only too well. Thus, Harrison et al. (1993) derive a generalised result in a static, perfect competition context which they call the 'home price effect' and the 'trade tax revenue effect' using vectors of endowments, of specific trade taxes, of world prices and of net imports of the country concerned. The survey by Baldwin and Venables (1995) applies a somewhat analogous approach, but they employ it to derive in a highly general fashion three groups of Single Market effects: three welfare effects in models of perfect competition, three arising in models of imperfect competition and scale, and finally an accumulation term (a higher equilibrium capital stock). The authors start from a general utility function for the economy, which includes vectors of border prices, of trade costs (including the tariff equivalent of import barriers) and of the number of product varieties, plus a scalar representing total spending on consumption. The static welfare effects are: the trade volume effect (linked to the price wedges created by barriers), a trade cost effect and a terms of trade effect. The effects emerging from models with scale and imperfect competition include: the output effect (change in output of sectors having prices different from average costs), a scale effect (value of changes of average costs in case of scale effects) and variety effects (impact on the number of differentiated products). Finally, the accumulation

effect (see also Baldwin 1989) is positive if the ratio of social return to social discounting is larger than one. As a categorisation of types of internal markets effects it is helpful. For example, on checking whether all the mentioned effects are represented in the literature, one finds that the gains of variety had not been analysed empirically for the Single Market until a contribution by Mohler and Seitz (2010). They are able to show that the gains of variety from intra-EU trade are traceable in the EU and matter in particular for 'newer' and smaller Member States in the period 1999–2008: for the smaller countries they range from 1% to 2.5% of GNP.

Price Convergence

With respect to *price convergence*, the Monti report shows (based on national price indices) that there was a general trend of price convergence in the EU-12 over the period 1980–1993. In consumer products, this trend has actually accelerated following EC1992. As expected, the trend of price convergence is sharpest for goods in highly traded sectors, whilst services remain characterised by significant price dispersion, presumably because their tradeability is low. The remaining price disparities (in goods) can be explained by quality aspects, remaining barriers, high levels of concentration in domestic markets in the EU (which enable price discrimination) and to some degree by indirect taxes. The nature of sectoral goods markets, such as homogeneous versus vertically differentiated goods, also matters. In Ilzkovitz et al. (2007), a similar analysis is presented based on the coefficient of variation of comparative price levels (not indices). The overall conclusion is an accelerated price convergence since EC1992 until 2005. Whereas the coefficient dropped from 20% in 1991 to 13% in 2005 for the EU15, it dropped from 39% to 26% in what are now the new EU Member States. The attribution to the Single Market is not easy since, for the EU15 (or the euro zone at least), strict monetary policy by the ECB and the higher price transparency caused by the euro have induced inflation levels to come down. There is a downward bias, too, since consumer goods imported from the world have often become cheaper. For the new Member States, the market discipline of EU accession has to be set against the Balassa–Samuelson effect of rapid catch-up growth, leading to sharper price increases, and quality upgrades in Central Europe, yielding a convergence upwards towards the EU25 average. Engel and Rogers (2004), based on consumer prices in a range of European cities, find that price dispersion falls considerably during the 1990s (confirmed by Hill 2004) and hardly or not at all in the period after the introduction of the euro (1 January 1999). The authors control for income, VAT rates and local labour costs and also employ a special model for dynamic price adjustments. Interestingly, although price disparities in non-tradeables, essentially services, remain larger than for goods in 2003, their decline in dispersion since 1990 is much sharper. Since, besides transport services and financial services, only intra-EU business services grew exceptionally fast in this period, such a convergence may follow from the close association of business services with manufactured output, in turn linked to increased intra-EU FDI.

It is instructive to complement the broad picture of price convergence in the Single Market with two telling sectoral examples: cars and telecom services. In cars, Goldberg and Verboven (2005) find strong evidence of convergence towards both the absolute and the relative version of the law of one price, with firmer evidence of the latter. The EU car market has remained notoriously difficult to integrate over time, first because of initially strong preferences for national cars in the bigger EU countries (Hocking 1980), amplified by an array of technical and fiscal barriers as well as explicit market segmentation via selective distribution systems. Goldberg and Verboven (2005), succeeded in showing that the long battle by the Commission to break down all these barriers, including a significant tightening of the exemption for car dealers from the normal competition rules on selective distribution, is reflected in a gradual but fairly sustained price convergence in both absolute and relative terms (given homogeneous definitions of cars across Member States). The opposite may be noted in telecoms

services. Pelkmans and Renda (2011) show that there is no such thing as an eCommunications Single Market, whether one studies the prices of 11 important telecoms services in the EU27 or the EU rules and institutional framework for that market. What is striking is the great emphasis on liberalisation of *national* telecoms markets in the EU, based on very similar rules and similar details to the (national) application of competition policy to this market. The upshot looks much like 27 liberalised national markets, not a Single Market. As a consequence, when defining a bilateral price disparity of 50% as a ratio of 150, the highest/lowest price ratios for the 11 services move in ranges from over 300 to over 4000, with the second highest/lowest ratios still from over 200 to beyond 2000.

How Integrated the Is Single Market: Can 'Home Bias' Tell Us More?

Market integration, traditionally measured by price convergence and/or trade flows (and to some extent, FDI stocks and people or worker mobility), can also be measured in a very different fashion: a decreasing 'home bias'. Home bias is based on a pure integration benchmark: in a truly integrated single market, economic agents from EU country A would assume an EU-wide outlook in all economic decisions, and hence purchase a share of intermediate and final goods domestically reflecting the relative size of the country in the EU economy; the remainder would be bought in the rest of the EU. One can correct such a benchmark for reasons of a common language (lowering trade costs) between two or more EU countries, contiguity between two or more EU countries (greater familiarity and better information), or business networks (see Combes et al. 2005) and other factors (such as sectoral differences: Chen 2004). Nevertheless, it is shown that 'home bias', defined as an excess of domestic purchases over the benchmark share, is quite high, an indication that market integration is not so deep. Even without customs, intra-EU frontiers still matter for economic conduct. Home bias indicators differ between authors but are invariably high. Thus, Head and Mayer (2000) find a corrected home bias of 25 in the mid-1970s and Nitsch (2000) one of 13 in 1979. But over time they all fall significantly (see also Delgado 2006); For example, Head and Mayer (2000) find 13 in 1995. Thus it would seem that deepening of market integration is reflected in falling home bias. In a comparison with the US Single Market (with pooled data for both and a unique equation for estimation), it turns out that EU countries' home bias is some 3–4 times that of US states (Pacchioli 2011). The key problem in interpreting home bias is that there is no underlying economic theory. Explanations for size and trends of bias remain somewhat arbitrary.

Productivity Effects of the Single Market

A whole literature has sprung up on the productivity and growth effects of the Single Market. In a micro-economic perspective, X-inefficiencies (or managerial slack facilitated by a lack of competitive pressure – see Schmidt (1997) for the state of the art) will be reduced, especially in the case of national incumbents in response to increased competition prompted by deepening of the Single Market. In a more aggregated perspective, Baldwin (1989) suggests a medium-term growth bonus in a simple Solow-type growth model arising from the productivity improvements induced by EC1992. Turning to empirics, not unlike some of the previous empirical contributions, 'the' Single Market is sometimes represented merely by 'the' EU as such. Notaro (2011) focuses primarily on the 'sensitive' sectors defined as expected to be hit by the removal of heavy cross-border barriers to competition. Using industry level panel data, he finds a positive productivity shock for those sectors of around 2% in 1992 and 1993. Bottasso and Sembenelli (2001), based on a panel of Italian firms, come to similar conclusions for Italy. Henrekson et al. (1997) had already found that the old 'customs-union-plus' (i.e. prior to 1985) had exercised a positive effect on EU economic growth. Halkos and Tzeremes (2009) find, not surprisingly, that the new EU Member States have enjoyed a higher growth effect from the

Eastern enlargement than the old Member States, after controlling for various factors and using different measurements. Both the Cecchini and Monti reports comprise a large number of sector studies showing the sectoral dynamics linked, often in detail, with the EC1992 programme. These studies leave no doubt that inter-sectoral reallocation and sectoral consolidation (and possibly, intra-firm improvements) amounted to a boost of technical efficiency and productivity. The extreme case of railway rolling stock, at first totally fragmented between EU countries irrespective of scale or R&D, is telling: of every four workers in 1987, no fewer than three were expected to be redundant once these markets could be subjected to open, competitive tendering, without even considering the impetus for technical progress in the EU railway sector.

The Single Market's Impact on Innovation

As discussed, the Single Market has undoubtedly fuelled competition in the EU. Such heightened rivalry may well strengthen the incentives for innovation in processes and/or products so as to enhance or protect their position in the market or indeed enter other national markets in the EU. This incentive mechanism operates via reduced price/cost margins. Yet another mechanism runs via new entry. In a rich theoretical approach, Vives (2008) finds that increasing market size increases cost reduction expenditure per firm while having ambiguous effects on the number of varieties offered. He also finds that decreasing the cost of entry increases the number of varieties but reduces cost reduction expenditure per variety. Relying on the pro-competitive effect of deepening market integration via EC1992, reducing price/cost margins (and hence average profitability) are observed by Griffith et al. (2006), in turn leading to an increase in R&D investment in manufacturing industry, subsequently translating into faster TFP (total factor productivity) growth. Note that this empirical contribution ignores services. Neither does it relate the Single Market effect on innovation to patents or other intellectual property rights. This means that the untapped potential for innovation effects is much larger still. Once the EU patent comes into being, one should expect a boost (*ceteris paribus*) in the number of patents given the incentives of large market size (all EU countries would always fall under such a patent) and much lower costs (Guellec and van Pottelsberghe 2007).

A complementary approach is to study the rate of diffusion/adoption of innovation, as undertaken by Surinach et al. (2011). Some 52% of firms in the EU rely on innovation adoption rather than self-generation of innovation. The authors employ a two-stage estimation procedure. First, they define the impact of the Single Market on both competition and cooperation of firms. Subsequently, they estimate how these 'channels' induce innovation adoption. The notion that much of EU innovation is actually jointly produced (in interfirm cooperation) and subsequently adopted as a result sits uneasily with the sole emphasis on competition in the literature quoted above. Competition (such as that due to the Single Market) is found to stimulate the adoption of process innovation from outside the firm, but not other forms of innovation. The authors obtain a trade-off: whilst more competitive pressure stimulates market exchanges of processes and technologies, it also tends to impede cooperation of firms, and hence joint innovation. The empirical link with productivity growth is found to be strongest for process innovation.

A Macro-Economic Perspective on Single Market Effects

Already, in the days of the EEC customs union, there had been attempts to come up with estimates of one-off GDP increments induced by EU market integration. The initially trivial gains found (see above) were suggested to be complemented by 'dynamic effects' (as they were called then), although strictly scale economies are analysed as comparative statics; much the same as for technical efficiency improvements. In later analyses higher one-off GDP increments were obtained for the internal market, ranging from a little over

1% up to the 4.5% in Emerson et al. (1988). These approaches are based on much more sophisticated models (e.g. scale and imperfect competition, and later CGE simulation with increasing sophistication, such as Allen et al. (1998), preceded by Harrison et al. (1996) and Gasiorek et al. (1992); see also Baldwin and Venables (1995)). Hoffmann (2000) makes an elaborate attempt to model the Single Market in the late 1990s as a partial one, looking carefully at issues of incomplete implementation. Using a variant of Harrison et al. (1996), he finds an increment of only 0.84% of EU GDP. In Straathof et al. (2008) the EU impact is investigated for a period of more than four decades. The authors are capable of showing that the (trade, FDI and GDP) gains from the Single Market differ over six periods they distinguish. They also demonstrate that the trade impacts of enlargements are higher than that of Single Market deepening. Straathof et al. (2008) first rely on gravity equations to identify the impact of the Single Market in bilateral trade in goods and services as well as FDI. The effects differ significantly between EU countries. Subsequently, the trade-enhancing effect of market integration (goods add mostly, services only a little, because data go until 2005) on GDP is estimated at some 2–3% but, due to expected reallocation, productivity improvements and innovation, the long-run impact for the EU would be nearly 10%.

Conclusions

This compact survey of the economic analysis of the impact of the Single Market shows a wealth of contributions. Analytical sophistication has greatly increased over time, both theoretically and empirically. Following the simple estimates of the effects of the EEC customs union on trade and GNP, more complicated models including several CGE variants and many applications of gravity approaches have been employed. Also macro-econometric models have been utilised in order to obtain macro-economic effects on GNP, price levels and employment. Some of this work has 'redone' customs union effects of the past or addressed the trade effects of various enlargements (which largely hinge on market access, too) but more and more the economic impact of a 'deep' Single Market' has become the central focus.

Nevertheless, the depth and wide coverage of the Single Market of today is hardly captured in these newer approaches – the temptation to reduce the internal market to goods only has remained very strong. The greatest challenge will be the better economic understanding of the EU Single Market in services, both for cross-border intra-EU trade and for establishment, hence, local provision. Moreover, the research agenda can be widened to comprise the economic impact of the EU patent, of the now 'deep' European capital market as well as of foreign direct investment and internal migration.

See Also

- Euro Zone Crisis 2010
- European Central Bank
- European Central Bank and Monetary Policy in the Euro Area
- European Cohesion Policy
- European Monetary Integration
- European Monetary Union
- European Union Budget
- European Union (EU) Trade Policy

Bibliography

Allen, C., M. Gasiorek, and A. Smith. 1998. The competition effects of the Single Market in Europe. *Economic Policy*, no. 27, October.

Anderson, J., and E. van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93: 170–192.

Baldwin, R. 1989. The growth effects of EC1992. *Economic Policy*, no. 9.

Baldwin, R., and A. Venables. 1995. Regional economic integration. In *Handbook of international economics*, ed. G. Grossman and K. Rogoff, vol. 3. Amsterdam/New York: Elsevier.

Baldwin, R., J. Francois, and R. Portes. 1997. The costs and benefits of enlargement: the impact on the EU and Central Europe. *Economic Policy*, no. 24, October.

Bottasso, A., and A. Sembenelli. 2001. Market power, productivity and the EU Single Market program: Evidence from a panel of Italian firms. *European Economic Review* 45: 167–186.

Breuss, F. 2002. Benefits and dangers of EU enlargement. *Empirica* 29(3): 245–274.

Breuss, F. 2009. An evaluation of the EU's fifth enlargement with special focus on Bulgaria and Romania, European Economy – *Economic Papers* 361, European Commission.

Chen, N. 2004. Intra-national vs. international trade in the EU: Why do national borders matter? *Journal of International Economics* 63(1): 93–11.

Combes, P., M. Lafourcade, and T. Mayer. 2005. The trade-creating effects of business and social networks: Evidence from France. *Journal of International Economics* 66(1): 1–29.

De Bruijn, R., H. Kox, and A. Lejour. 2008. Economic benefits of an integrated European market for services. *Journal of Policy Modeling* 30(2): 301–319.

Delgado, J. 2006. Single Market trails home bias. *Bruegel Policy Brief* 2006/05. http://www.bruegel.org/

Egger, P., and M. Larch. 2011. An assessment of the Europe Agreements' effects on bilateral trade, GDP and welfare. *European Economic Review* 55(2): 263–279.

Emerson, M., et al. 1988. The economics of 1992. *European Economy*, no. 35, April.

Engel, Ch., and J. Rogers. 2004. European product market integration after the euro. *Economic Policy*, July, no. 19.

European Commission. 1988. *Research on the costs of non-Europe (20 volumes)*. Luxembourg: EC Publications Office, Documents Series.

European Commission. 1996. Economic valuation of the internal market. *European Economy*, Reports and studies, no. 4, December.

European Commission. 1997. *Single Market review (38 volumes)*. Luxembourg/London: Office of Official Publications of the EC/Kogan Page.

Gasiorek, M., A. Smith, and A. Venables. 1992. Trade and welfare: A general equilibrium model. In *Trade flows and trade policy after '1992'*, ed. L. Winters. Cambridge: Cambridge University Press.

Gasiorek, M., A. Smith, and A. Venables. 2002. The accession of the UK to the EC: A welfare analysis. *Journal of Common Market Studies* 40(3): 425–447.

Geroski, P. 1989. The choice between diversity and scale. In *1992 – Myths and Realities*, ed. E. Davis et al. London: Centre for Business Strategy, London Business School.

Goldberg, P., and F. Verboven. 2005. Market integration and convergence to the Law of One Price: Evidence from the European car market. *Journal of International Economics* 65: 49–73.

Griffith, R., R. Harrison, and H. Simpson. 2006. *Product market reform and innovation in the EU*. CEPR discussion paper, no. 5849, September.

Guellec, D., and B. van Pottelsberghe. 2007. *The economics of the European patent system*. Oxford: Oxford University Press.

Halkos, G., and N. Tzeremes. 2009. Economic efficiency and growth in EU enlargement. *Journal of Policy Modelling* 31: 847–862.

Harrison, G., T. Rutherford, and D. Tarr. 1996. Increased competition and completion of the market in the EU. *Journal of Economic Integration* 11(3): 332–365.

Harrison, G., T. Rutherford, and I. Wooton. 1993. An alternative welfare decomposition for customs unions. *Canadian Journal of Economics* 26(4): 961–968.

Head, R., and T. Mayer. 2000. Non-Europe: The magnitude and causes of market fragmentation in the EU. *Weltwirtschaftliches Archiv* 126: 2.

Henrekson, M., J. Torstensson, and R. Torstensson. 1997. Growth effects of European integration. *European Economic Review* 41(8): 1537–1557.

Hill, R. 2004. Constructing price indexes across space and time: The case of the European Union. *American Economic Review* 94(5): 1379–1410.

Hocking, R. 1980. Trade in motorcars between the major European producers. *Economic Journal* 90: 1. (September).

Hoffmann, A. 2000. The gains from partial completion of the Single Market. *Weltwirtschaftliches Archiv* 136(4): 601–629.

Ilzkovitz, F. et al. 2007. Steps towards a deeper economic integration: The internal market in the twenty first century. *Economic Papers* no. 271, January. European Commission, Brussels.

Johnson, H. 1958. The gains from freer trade in Europe, an estimate. *Manchester School*, 26.

Magee, C. 2008. New measures of trade creation and trade diversion. *Journal of International Economics* 75(2): 349–362.

Mayes, D. 1978. The effects of economic integration on trade. *Journal of Common Market Studies* XVII: 1.

Miller, M.H., and J.E. Spencer. 1977. The static economic effects of the UK joining the EEC: A general equilibrium approach. *Review of Economic Studies* 44(136): 71.

Mohler, L., and M. Seitz. 2010. *The gains from variety in the EU*. Muenchen discussion paper, no. 2010-24, March.

Mueller, J. 1981. Competitive performance and trade within the EEC: Generalisations from several case studies with specific reference to the West German economy. *Zeitschrift fuer die gesamte Staatswissenschaft* 137: 3.

Nitsch, V. 2000. National borders and international trade: Evidence from the EU. *Canadian Journal of Economics* 33: 1091–1105.

Notaro, G. 2011. European integration and productivity: Exploring the early effects of completing the internal market. *Journal of Common Market Studies* 49: 845–869.

Pacchioli, C. 2011. Is the EU internal market suffering from an integration deficit? *CEPS Working Document*, no. 348, May. CEPS, Brussels. http://www.ceps.eu/

Panagariya, A. 2000. Preferential trade liberalization: The traditional theory and new developments. *Journal of Economic literature* 38(2): 287–331.

Pelkmans, J. 2011. European Union Single Market: Design and development. In *The New Palgrave dictionary of economics*, forthcoming.

Pelkmans, J., and A. Renda. 2011. Single eComms market? No such thing. . .. *Communications & Strategies* 82: 21–42.
Petith, H.C. 1977. European integration and the terms of trade. *Economic Journal* 87: 262–272.
Prewo, W. 1974. Integration effects in the EEC: An attempt at quantification in a general equilibrium framework. *European Economic Review* 3.
Scaperlanda, A. 1967. The EEC and US foreign investment: Some empirical evidence. *Economic Journal* 77: 22–26.
Schmidt, K.M. 1997. Managerial incentives and product market competition. *Review of Economic Studies* 64(2): 191–213.
Scitovsky, T. 1958. *Economic theory and Western European integration*. London: Allen & Unwin.
Sorensen, P.B. 1998. Discussion. *Economic Policy*, no. 27, October.
Straathof, B., G.J. Linders, A. Lejour, and J. Moehlmann. 2008. The internal market and the Dutch economy – implications for trade and economic growth. *CPB Document* no. 168, September. http://www.cpb.nl/
Surinach, J., F. Manca, and R. Moreno. 2011. Extension of the study on the diffusion of innovation in the internal market, European Commission, DG Ecfin. *Economic Papers* no. 438, February.
Verdoorn, P., and A. Schwartz. 1972. Two alternative estimates of the effects of EEC and EFTA on the pattern of trade. *European Economic Review* 3(3): 291–335.
Viaene, J.M. 1982. A customs union between Spain and the EEC: An attempt at quantification of the long-term effects in a general equilibrium framework. *European Economic Review* 18(2): 345–368.
Vives, X. 2008. Innovation and competitive pressure. *Journal of Industrial Economics* 61(3): 419–469.
Winters, L.A. 1987. Britain in Europe: A survey of quantitative trade studies. *Journal of Common Market Studies* 25: 315–335.

European Union's Common Agricultural Policy (CAP)

Alan Swinbank

Abstract

The EU's CAP has changed significantly over the years, largely as a result of international pressure through GATT and the WTO, but it still has the support of farm incomes as its main concern. The old CAP of market price support has more or less been displaced by decoupled income support, in the shape of the Single Payment Scheme. Cross compliance applies, addressing concerns about European agriculture's multifunctionality; and there is a renewed anxiety about food security. The CAP's Second Pillar, with targeted support for rural development and environmental protection, plays a subsidiary role.

Keywords

Agriculture; CAP; Common Agricultural Policy; Decoupled payments; EU; Farm income; Food security; General Agreement on Tariffs and Trade; GATT; Multifunctionality; Rural development; Second Pillar; World Trade Organization; WTO

JEL Classifications

F13; F15; Q18

Introduction

Many have almost believed the *Common Agricultural Policy* (CAP) and the *European Union* (EU) to be synonymous terms. For some, the CAP was an essential building-block of the EU; for others, the perceived economic failures of the policy condemned the whole European endeavour.

From the outset, the Treaty establishing the European Economic Community (the earlier incarnation of today's EU) declared that the common market should 'extend to agriculture and trade in agricultural products', and that one of the activities of the EEC would be 'the adoption of a common policy in the sphere of agriculture' (Articles 38 and 3(d)). Neither the evolution of the EEC into today's EU, following various treaty changes, nor the respective competencies of the EU's institutions – (European) Commission, Council of Ministers, European Council, European Parliament etc. – are discussed here (but see Nugent 2010).

The form that the CAP would take was unclear, but various objectives were laid out in Article 39. These were:

(a) 'to increase agricultural productivity by promoting technical progress and by ensuring the rational development of agricultural production and the optimum utilisation of the factors of production, in particular labour;
(b) thus to ensure a fair standard of living for the agricultural community, in particular by increasing the individual earnings of persons engaged in agriculture;
(c) to stabilise markets;
(d) to assure the availability of supplies;
(e) to ensure that supplies reach consumers at reasonable prices'.

Thus the Treaty set an income objective; although quite who was thought to be included in the 'agricultural community', or 'engaged in agriculture' was never entirely obvious. Despite its lack of clarity, this 'farm income' objective has remained a distinctive feature of the CAP ever since; and differentiates 'farming' from all other economic activities in the EU.

The CAP has changed *substantially*. Other policy objectives such as food safety, animal welfare, and environmental protection have come to the fore, and concerns about food security resurfaced in the late 2000s; but the obsession with 'farm income' has remained throughout. This text first sets out the mechanisms that were put in place in the 1960s, which might be characterized as the *Old CAP* of market price support. Then it explains the policy changes of the early 1990s, which began a decoupling of income support, leading to the establishment of the Single Payment Scheme (SPS) in the 2000s. The fourth section introduces the concept of *multifunctionality*, whilst the fifth outlines the *Second Pillar* of the CAP (the First Pillar being market and income support). At the time of writing the EU is engaged in a review of the role the CAP should play after 2013, and the final section sets out some of the issues raised.

EU legislation impacting the farm, food and rural sectors extends beyond the confines of the CAP, as outlined here. Food safety and animal and plant health regulations are necessary to support the operation of the single market, and its Renewable Energy Directive provides incentives for biomass to be used for energy, particularly biofuels in transport, for example.

The Old CAP: Market Price Support

The creation of the CAP was a protracted, and often contested, process (Knudsen 2009), keenly observed by the EEC's trading partners, but the result in the mid-1960s was a CAP that involved extensive regulation of the markets for farm products and processed foods. This policy dominated CAP budget expenditure, and perceptions of the CAP, until policy reforms were initiated in the early 1990s.

Farm-gate prices were often two or three times world market prices, and the policy was premised on the idea that, by raising farm *revenues*, farm *incomes* would also increase. This ignored three important limitations. First, given the heterogeneity of European agriculture, and in particular the variability in the size of farm businesses, price support had little absolute impact on the revenues of small farms, but considerably influenced those of larger operators. It was not until the early 1990s that the Commission (1991) claimed that 80% of CAP budget support benefited 20% of farms, accounting for 'the greater part of the land used in agriculture'.

Second, it confused revenues with incomes. An increased demand for farm products leads to an increase in the derived demand for inputs, boosting the sales of fertilizer and machinery suppliers, benefiting landowners rather than tenant farmers, and retaining in farming marginal producers who might otherwise have quit, thus perpetuating the farm 'problem'.

Third policy-makers neglected the market regulating effects of the price mechanism and, as productivity improvements fuelled the ability of European farmers to increase supply, surpluses soared, leading to the butter mountains and wine lakes so often associated with the CAP.

Details and terminology differed from one commodity to another, but most had three key policy mechanisms in place (Harris et al. 1983). First, high import taxes were applied. Prior to the conclusion of the Uruguay Round of GATT

(General Agreement on Tariffs and Trade) negotiations, and the new WTO (World Trade Organization) system of international trade regulation that applied from 1995, these often took the form of *variable import levies* that were regularly expanded or contracted to plug the gap between the higher, and EU-determined, minimum import price (often known as a *threshold* price), and the lower, and variable, world market price. These import barriers were converted into fixed tariff equivalents (*tariffication*), and reduced; but for many products (e.g. sugar) they remained prohibitively high. A successful end to the WTO's Doha Round would bring substantial tariff reductions, of up to 70% in some instances (Daugbjerg and Swinbank 2009, pp. 54, 168).

The internal market price was directly supported: farmers (or more usually traders, or manufacturers of first-stage processed products, such as butter) could sell product to the intervention agencies at fixed prices: whence the butter mountains. Once acquired, the product might be sold back onto the EU market if prices strengthened, sold into lower-priced outlets (for example skim milk powder was used for animal feed), or at reduced prices on world markets. Over the years intervention prices have been reduced substantially, not just in real terms but also in nominal terms, and by the beginning of the 2010s intervention played only a limited role in CAP market price support.

The third element in the tripod of support was an export subsidy (known as an *export refund*). Traders could claim this export subsidy if product was sold outside the EU. As production expanded at a faster rate than consumption, the result was, first, a squeezing-out of imports, and then a growing volume of subsidized exports, depressing world market prices and angering other nations that saw their trading interests threatened. It was an anomaly that the GATT allowed export subsidies on agricultural products, but not on manufactured goods. The Uruguay Round agreements imposed more disciplines on the use of export subsidies on agricultural products, whilst leaving the EU largely untroubled; but successive CAP reforms have significantly reduced the EU's reliance on export subsidies, and they will be eliminated from the world trade regime if the Doha Round is concluded.

Simple welfare economic critiques of the CAP, showing that the gains to producers were outweighed by the losses to consumers and taxpayers, were repeated many times, but without a notable impact on policy. This led to the conclusion that the analysis was incomplete for one, or both, of two reasons. First, that political forces were too entrenched to be readily overcome, leading to political economy appraisals of the CAP (e.g. Senior Nello 1984), and attracting the interest of political scientists. Second, that an undue emphasis on perfectively competitive markets overlooked the prevalence of market failure: policy-makers were perhaps acting rationally in their design and defence of the CAP.

A Changed Regime: The Newer CAP of Decoupled Income Support

The old CAP survived the attempts of Sicco Mansholt, the first Commissioner for Agriculture, to reform the policy in 1968, and the accession of the United Kingdom (allegedly antagonistic to the CAP) in 1973. Nonetheless, some modest changes were introduced to try to limit the growth in surpluses and budget expenditure without recourse to significant price cuts, such as milk quotas in 1984.

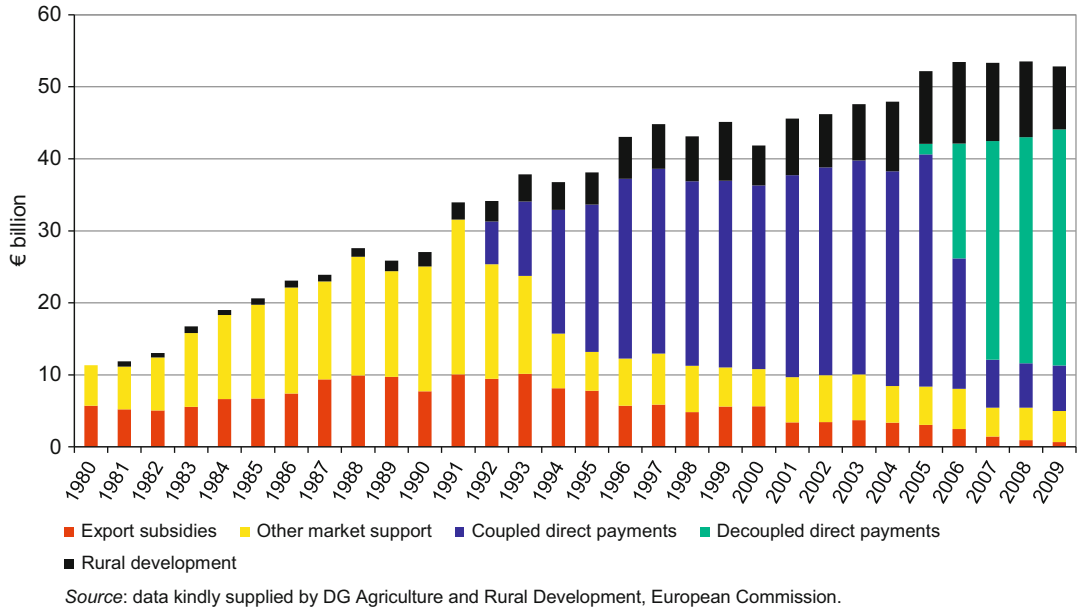
It had long been assumed that the soaring budget cost of the CAP would breach the EU's budget ceilings, forcing CAP reform. Some authors suggest that this was a major factor in the 1992 reform, whereas others focus upon the constraints imposed by the Uruguay Round of GATT negotiations (Cunha with Swinbank 2011, Ch. 5). All analysts, however, are agreed that this was a political process, in which the rent-seeking interests of the farm lobby were overwhelmed by competing forces, either in a zero-sum game in the allocation of EU budget funds, or as a necessary trade-off to secure other objectives in the Uruguay Round, rather than a considered response to an economic critique.

The exceptional treatment that agriculture had received in the GATT was largely a product of the

exigencies of US farm policy in the GATT's formative years (Josling et al. 1996, Chs. 1 and 2). By the mid-1980s circumstances had changed, and the USA and others were determined to negotiate new disciplines governing farm support in the Uruguay Round negotiations, launched in Punta del Este, Uruguay, in 1986, and finally concluded in a formal ceremony in Marrakesh, Morocco, in 1994 (Croome 1999).

As well as curbs on the use of export subsidies on agricultural goods, and greater market access opportunities, the GATT negotiators set out to *decouple* farm support. If farm policies acted to influence domestic production, and hence a country's net trade position, this impinged on the trading interests of other nations, and was a legitimate GATT concern. Consequently the Uruguay Round Agreement on Agriculture (URAA) slotted farm support policies into one of three broad categories, the so-called *amber*, *blue* and *green boxes*. Policies that *did* distort trade (the amber box) would be subject to new WTO disciplines. Policies deemed to have minimal impact on production would be exempt from subsidy constraints, and fell within the green box (Annex 2 of the URAA) (Daugbjerg and Swinbank 2009, p. 54). This specified that 'decoupled income support' had to satisfy both *ex ante* and *ex post* criteria: policy design must match a set of specifications (that 'payments in any given year shall not be related to, or based on, the type or volume of production. . . undertaken by the producer in any year after the base period', for example), and the over-arching 'fundamental requirement that they have no, or at most minimal, trade-distorting effects or effects on production'.

The EU 'decoupled' the bulk of its farm support in two steps. The MacSharry reforms of 1992 – driven by the then Agriculture Commissioner Ray MacSharry –achieved a partial decoupling. Cereal and beef prices were reduced, and farmers were compensated for the implied revenue loss with area payments on land sown, and in enhanced (headage) payments on beef animals kept. Second, starting with Franz Fischler's reforms of 2003, and subsequently extended to other sectors, these area and headage payments were converted into the more-decoupled SPS (Cunha with Swinbank 2011).



Source: data kindly supplied by DG Agriculture and Rural Development, European Commission.

European Union's Common Agricultural Policy (CAP), Fig. 1 The changing structure of CAP budget expenditure 1980–2009, € billion, current prices

The scale of the policy change can be gauged from Fig. 1. Until the 1990s expenditure on export subsidies and other market price support mechanisms dominated, with some expenditure on rural development (now dubbed the *Second Pillar* of the CAP, and discussed further below). With the MacSharry reforms of 1992 direct payments to support farm incomes began to displace market price support; and the Fischler reforms of 2003, and subsequently, switched these into the decoupled SPS, further squeezing, but not quite displacing, the old CAP of market price support. The relative budgetary importance of the three strands of today's CAP can be gauged by the budget allocations for 2011: some h39.8 billion for direct aids (mainly the SPS), with a mere h3 billion allotted to interventions in agricultural markets (the old CAP); and h12.6 billion for rural development (European Commission 2011a). Collectively the CAP still accounts for almost 44% of expected budget payments.

The area and headage payments of the MacSharry reforms were only partially decoupled – crops had to be sown and animals kept – and so did not qualify as green box payments under the URAA. Instead they were classified in an intermediate category, the *blue box*, which like the green box was not subject to expenditure limits, ensuring that the MacSharry reform fitted the new URAA. This was seen by many of the EU's trading partners as a temporary relaxation of the subsidy rules to be swept away in the next (Doha) round of negotiations. In its periodic declarations to the WTO the EU has claimed that the SPS has switched the bulk of its income support into the green box, meeting any new subsidy constraints that the Doha Round might bring.

Despite the claim to be a *common* agricultural policy, there are a number of differences in the way the SPS is applied. One relates to it operating on an historic or a regionalized basis, with Member States (and sometimes regions within them) opting for one or the other.

The original idea was that farm businesses would be entitled to SPS payments based on their claims for area and headage payments in a base period. If the annual claim had averaged h*a* per year, and *b* hectares had been used to justify this claim, then they would receive an SPS allocation of *b* entitlement hectares at a payment rate of h*a*/*b* per hectare. This could be claimed against on an annual basis, provided *b* hectares of farmland were kept in good agricultural and environmental condition (Daugbjerg and Swinbank 2009, p. 136). This meant that not all farmland had an SPS entitlement associated with it – because some farm enterprises, for example outdoor pigs, did not give rise to an entitlement – *and* that per hectare SPS payments varied significantly, depending on the farm's particular enterprise mix in the base period.

Alternatively, Member States were entitled to pool the SPS entitlements for a region, and make flat-rate payments on all eligible land in that region. Compared to the historic model of the previous paragraph, some farms lost high per-hectare payments that they might otherwise have received, whilst others gained. Moreover, more land was brought into the scheme, extending the scope of *cross-compliance* – respect of statutory standards, and good agricultural and environmental practice – which was required for participation.

Although the original intent, back in 1992, was to *compensate* farmers for the cut in intervention prices then experienced, the SPS is now described as 'an income support for farmers' (Article 1, Regulation 1782/2003). This posed a dilemma for the EU contemplating enlargement to embrace (in 2004 and 2007) 10 countries from Central and Eastern Europe (the CEECs). If seen as compensation for past price cuts, it would be illogical to make the same payments to farmers who had not experienced those reductions. Moreover, they might well distort competition if paid in part of the enlarged EU but not throughout, even though justified in the WTO as decoupled payments. Logic suggested that *compensation* should be temporary, and that payments should be phased out. If, however, they were seen as a permanent form of CAP *income* support, with the potential to distort competition, they could not be denied to the new entrants: an incongruous outcome given the EU's declared objective of securing economic and social cohesion.

The debate split the member states, with some (such as France) arguing for an extension of the full CAP to the new members, and others (such as the United Kingdom) arguing for CAP reform, as on previous and subsequent occasions. Eventually the SPS was extended to the new member states (in most in a slightly different format known as the Single Area Payment Scheme, or SAPS); albeit with a phased introduction to lessen the initial impact on the EU's budget, and any disruptions that a sudden inflow of wealth into rural areas might have. EU funding was at 25% of the EU15 rate in the first year, rising to 100% in 2013 (2016 in Bulgaria and Romania), with the possibility of nationally funded 'top-ups' (European Commission 2004). The phased introduction of the payments, and the initial determination of each new member state's basic budgetary entitlement, has been a source of dissatisfaction ever since.

Justifying the continuing need for income support, the European Commission (2010a, p. 5) still claims that agricultural income is 'significantly lower. . . than in the rest of the economy'; which is rather at odds with critics such as Hill (2000, p. 346) who, writing before Eastern enlargement, considered there to be 'considerable evidence' that 'there is no income problem for many farm families, in the sense that their overall incomes are not low but compare favourably with households in general'. Payments are very unequally spread, with no indication that they are targeted at individuals in need. In 2009 for example, in France, 21% of recipients of all direct aids received h2,000 or less, and were collectively paid less than 1% of the funds disbursed, whereas 1.5% of claims were for h100k or more, accounting for over 10% of the monies paid out (European Commission 2011b, Annex 4.1). Commission proposals to cap the annual payment – at, say, h300k per farm – have as yet been rejected by the Council of Ministers.

Multifunctionality

Farming generates both positive and negative externalities; and in the densely populated Old World, where fauna, flora, landscape and culture have been moulded by thousands of years of land use, and few wilderness areas remain, there are concerns that policy change could lead to an unwelcome diminution in positive, and exacerbation of negative, externalities. The SPS, it has been claimed, provides 'support for basic public goods desired by European society' (European Commission 2010a, p. 4).

As with other economic activities, agriculture is subject to a number of legislative provisions that are designed to internalize or curb its negative externalities. Thus, in nitrate vulnerable zones, restrictions are placed on the spreading of animal slurry; and the Welfare of Laying Hens Directive governs caged systems of egg production. But, unlike other industries, and challenging the polluterpays principle and the concept of comparative advantage, the CAP often seeks to compensate additional costs. Many of these environmental requirements are built into the cross-compliance provisions of the SPS.

The concept of *multifunctionality* adds an extra dimension. Here the focus is on positive externalities *jointly* produced with marketable farm outputs in traditional farming systems, which – it is feared – would not be produced if farming practices changed. The European Commission no longer uses the term – although some Member States do – but it was a concept the EU was pushing before the outset of the Doha Round. Perhaps the high point in the Commission's defence of multifunctionality came when it declared: 'The fundamental difference between the European model and that of our major competitors lies in the multifunctional nature of Europe's agriculture and the part it plays in the economy and the environment, in society and in preserving the landscape, whence the need to maintain farming throughout Europe and to safeguard farmers' incomes' (Commission 1998, p. 8). But all this rather assumed that: (1) the counterfactual was a less-preferred outcome; (2) these valued externalities could *only* be supplied if traditional farming practices were preserved; and (3) farmers would continue farming in a traditional way if their incomes were safeguarded; delivering (4) the multifunctional attributes for which European citizens were presumed willing to pay.

The Second Pillar

In addition to the mandatory and altruistic actions of farmers, delivering positive externalities associated with 'diverse forms of agriculture, rich in tradition' (Commission 1998, p. 7), farms are eligible under Pillar II for funding for environmental and rural development schemes. Unlike Pillar I, which is fully funded through the EU budget, member states *co-finance* Pillar II expenditure. The Second Pillar, like the first, has evolved over the years (Thomson et al. 2010, pp. 378–83), and its present incarnation is the Rural Development Regulation (No 1698/2005) covering the period 2007–13. This offers a menu of policy options that Member States can implement, set out in what are referred to as three thematic axes, together with the *Leader* (Liaison Entre Actions de Développement de l'Économie Rurale) programme.

The first axis focuses on the competitiveness of the farm sector, and dates back to the early years of the CAP. It provides for a wide array of on- and off-farm investments, including farm modernization, early retirement, and transitional support for producer groups, for example. The second axis, entitled 'improving the environment and the countryside', has its origins in the Less-favoured Area (LFA) directive of 1975, authorizing additional support in the LFAs, which by 2005 covered 54% of the EU27's utilizable agricultural area (European Commission 2010b, p. 86). LFA payments are still a major form of expenditure under this heading, but other schemes include establishing forests, and agri-environment payments to farmers (and sometimes other land managers) who commit to supplying environmental benefits in addition to those required by cross-compliance under Pillar I.

Axis 3 is concerned with the 'quality of life in rural areas and diversification of the rural economy', with a slighter wider remit than farm households, fostering for example 'village renewal and development'. Leader is a 'bottom-up' approach, allowing for local initiatives to design and implement development programmes across the three thematic axes. A Leader project in the UK, for example, helped in the rejuvenation of reed and sedge harvesting on the Norfolk Broads, a defining characteristic of that local environment (European Commission 2006, p. 21). It is not entirely clear why Axis 3 and Leader are funded through the CAP, rather than forming part of the EU's wider regional and cohesion policies.

There is considerable diversity in the way the Member States implement the Regulation. Their Rural Development Plans for the period 2007–13 are supposed to devote at least 10% of their EU funds to Axis 1, 25% to Axis 2 and 10% to Axis 3. Across the EU, Axis 2 is the most popular at 47%, followed by Axis 1 (reflecting traditional CAP concerns) at 33%, whilst Axis 3 (rural development) at 17% is the least (with Leader funds allocated to the three axes where appropriate). Member states have expressed significantly different preferences: Ireland for example allocates the minimum amount allowed to Axis 3, *all* channelled through Leader, whereas 80% of its funds are spent in Axis 2 (European Commission 2010b, p. 139–40).

The Post-2013 CAP?

The future CAP is closely linked with the EU's *Financial Perspective* for 2014–20. When the current Financial Perspective (for 2007–13) was thrashed out in December 2005, a 'full, wide-ranging review' was promised, 'covering all aspects of EU spending, including the Common Agricultural Policy', which would input into the discussions on the post-2013 Financial Perspective. That debate has been under way for some time (Cunha with Swinbank 2011, p. 183, 195–200). The Commission's initial ideas for the post-2013 CAP were published in November 2010 (European Commission 2010a); but it will be 2012, or even later, before any new legislation is in place. The role that the European Parliament will play, given its enhanced powers over the CAP following ratification of the Treaty of Lisbon, has yet to be experienced. In particular it now shares legislative responsibility with the Council over major CAP decisions – a procedure that used to be known as *co-decision* – in contrast to its earlier simple consultative role; and it has more control

over CAP funding, as the old concept of *compulsory expenditure* – elements of the CAP budget that Parliament could not touch – has lapsed (Cunha with Swinbank 2011, p. 40–2).

A number of themes emerge from the Commission's November 2010 communication. First, as mentioned earlier, it suggests that the farm income problem persists, despite more than forty years of CAP support; and that a continuation of the SPS (or something similar) is needed, albeit with a more 'equitable and balanced' distribution between Member States. Second, that there should be a further 'greening' of support, to enhance the 'environmental performance of the CAP'.

A third theme is the suggestion that a 'strong' CAP is necessary to 'guarantee long-term food security for European citizens' (p. 2), reflecting public concerns about population growth, rising dietary aspirations in China and India in particular, the use of arable land for biofuels, and global warming. Thus it is too soon to say whether the progressive liberalization of the CAP, launched in the early 1990s, will be sustained, or instead there will be retrenchment, and a revival of elements of the old CAP, in the belief that this enhances the food security of Europe's citizens.

See Also

- ▶ Euro Zone Crisis 2010
- ▶ European Central Bank
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Cohesion Policy
- ▶ European Labour Markets
- ▶ European Monetary Union
- ▶ World Trade Organization

Bibliography

Commission of the European Communities. 1991. *The development and future of the CAP*, Reflections Paper of the Commission. COM(91)100. Brussels: Commission of the European Communities.

Commission of the European Communities. 1998. *Proposals for Council regulations (EC) concerning the reform of the common agricultural policy*, COM (1998)158. Brussels: Commission of the European Communities.

Croome, J. 1999. *Reshaping the world trading system: A history of the Uruguay Round*. The Hague: Kluwer Law International.

Cunha, A. with Swinbank, A. 2011. *An inside view of the CAP reform process: Explaining the MacSharry, Agenda 2000, and Fischler reforms*. Oxford: Oxford University Press.

Daugbjerg, C., and A. Swinbank. 2009. *Ideas, institutions and trade: The WTO and the curious role of EU farm policy in trade liberalization*. Oxford: Oxford University Press.

European Commission. 2004. *Chapter 7 – Agriculture. Enlargement Archives*. http://ec.europa.eu/enlargement/archives/enlargement_process/future_prospects/negotiations/eu10_bulgaria_romania/chapters/chap_7_en.htm. Accessed 24 Feb 2011.

European Commission. 2006. *The leader approach – A basic guide*. Luxembourg: Office for Official Publications of the European Communities.

European Commission. 2010a. *The CAP towards 2010: Meeting the food, natural resources and territorial challenges of the future*, COM(2010)672. Brussels: European Commission.

European Commission. 2010b. *Rural development in the European Union statistical and economic information report 2010*. Brussels: European Commission.

European Commission. 2011a. *2011 Budget online*, Title 0.5. http://eur-lex.europa.eu/budget/data/LBL2011/EN/SEC03.pdf. Accessed 9 Feb 2011.

European Commission. 2011b. *Indicative figures on the distribution of aid, by size-class of aid, received in the context of direct aid paid to the producers according to Council Regulation (EC) No 1782/2003 and Council Regulation (EC) No 73/2009 (financial year 2009)*. http://ec.europa.eu/agriculture/fin/directaid/2009/annex1_en.pdf. Accessed 7 Feb 2011.

Harris, S., A. Swinbank, and G. Wilkinson. 1983. *The food and farm policies of the European community*. Chichester: Wiley.

Hill, B. 2000. *Farm incomes, wealth and agricultural policy*. 3rd ed. Aldershot: Ashgate.

Josling, T.E., S. Tangermann, and T.K. Warley. 1996. *Agriculture in the GATT*. Basingstoke: Macmillan.

Knudsen, A.-C.L. 2009. *Farmers on welfare: The making of Europe's common agricultural policy*. London: Cornell University Press.

Nugent, N. 2010. *The government and politics of the European Union*. 7th ed. Basingstoke: Palgrave-Macmillan.

Senior Nello, S. 1984. An application of public choice theory to the question of CAP reform. *European Review of Agricultural Economics* 11: 261–283.

Thomson, K., P. Berkhout, and A. Constantinou. 2010. Balancing between structural and rural policy. In *EU policy for agriculture, food and rural areas*, ed. A. Oskam, G. Meester, and H. Silvis. Wageningen: Wageningen Academic Publishers.

Evans, Griffith Conrad (1887–1973)

Herbert A. Simon

Keywords

Calculus of variations; Comparative statics; Evans, G. C.; Mathematical economics; Non-equilibrium dynamics; Two-sector models

JEL Classifications

B31

A distinguished American mathematician and pioneer mathematical economist, Evans was born on 11 May 1887 in Boston, Massachusetts. Educated in mathematics at Harvard University (AB, 1907; MA, 1908; Ph.D., 1910), he spent two years as a postdoctoral fellow studying with Vito Volterra at the University of Rome, then joined the faculty of the Rice Institute in Houston, Texas, where he taught from 1912 to 1934. In 1934, he became chairman of the mathematics department at the University of California, Berkeley, retaining that position until his retirement in 1954. He died on 8 December 1973, at the age of 86.

Evans's important contributions to mathematics, especially in functional analysis and potential theory, earned him membership in the National Academy of Sciences in 1933, as well as numerous other professional honours. His interest in mathematical economics became evident about 1920, when he gave his first series of lectures on that subject at the Rice Institute, and it continued up to the time of his retirement, his last publication on the subject appearing in 1954. It is likely that his initial contact with mathematical economics took place in Italy and France, for he shows great familiarity with the work of such writers as Pareto, Amoroso and Divisia, who were flourishing during and after his early Continental sojourn. Among earlier writers in mathematical economics, he mainly cities Cournot and Jevons; among his contemporaries, Irving Fisher, Henry Schultz, and Henry Moore.

Evans's most important work in economics is his *Mathematical Introduction to Economics* (1930), which also contains materials from his earlier papers. In the book and his other publications, he applied the calculus and the calculus of variations to problems of monopoly, duopoly and competition, and to a whole range of problems of comparative statics, including the incidence of taxes and the effects of tariffs. His approach was quite different from that of Walras (to whom he does not refer in his book), in that most of his models dealt with one or a few actors in a single market, or a small number of markets. In his 'Maximum Production Studied in a Simplified Economic System' (1934, p. 37), he gave clear expression to his attitude toward general equilibrium models: 'Large numbers of simultaneous equations in a large number of variables convey little information ... about an economic system.'

In Evans's models, supply was generally described in terms of a cost function rather than a production function. To deal with macroeconomic problems, he constructed aggregate models, and, in order to provide a rationale for such models, he and his students made a deep study of the problem of index number construction. Their starting point was the work of Irving Fisher and François Divisia.

Evans's books and his articles were an important resource for early American students with an appetite for mathematical economics, who prior to their publication found an extremely sparse literature on which to graze. Samuelson, for example, mentions Evans as one whose works he 'pored over' when working on the *Foundations of Economic Analysis*. Moreover, Evans's methods of modelling economic situations gave new impetus to the approach of comparative statics, especially in application to macroeconomic problems. He constructed an early (perhaps the first) two-sector aggregative model containing a consumption good and a capital good; and he saw the power of second-order conditions of stability in reasoning about comparative statics, thereby anticipating by more than a decade Samuelson's important contributions to that topic.

Evans did only a little work in non-equilibrium dynamics, although he saw clearly the need for

further development of that subject. In his 'Simple Theory of Economic Crises' (1931, p. 61), he complained that 'the fact of lack of equilibrium in economic systems continually, and practically, stares us in the face; yet the principal discussion from a theoretical point of view has been of equilibrium, and thus at one stroke has eliminated a major issue.'

Evans was a fellow of the Econometric Society, and one of its founders. His principal influence upon the progress of economics came through the methodologies employed in his book, and through the work of his students, among whom were Francis W. Dresch, Kenneth May, C.F. Roos and Ronald W. Shephard, and one step removed, Lawrence W. Klein and Herbert A. Simon, who were colleagues or pupils of these students.

Selected Works

1922. A simple theory of competition. *American Mathematical Monthly* 29: 371–380.
1924. The dynamics of monopoly. *American Mathematical Monthly* 31: 77–83.
1925a. Economics and the calculus of variations. *Proceedings of the National Academy of Sciences of the USA* 11: 90–95.
1925b. The mathematical theory of economics. *American Mathematical Monthly* 32: 104–110.
1929. Cournot on mathematical economics. *Bulletin of the American Mathematical Society* 35: 269–271.
1930. *Mathematical introduction to economics.* New York: McGraw-Hill.
1931. A simple theory of economic crises. *American Statistical Association Journal* 26- (Supplement): 61–68.
1932a. Stabilité et dynamique de la production dans l'économie politique. In *Mémorial des Sciences Mathématiques 61.* Paris: Gauthier-Villars.
1932b. The role of hypothesis in economic theory. *Science* 75: 321–324.
1934. Maximum production studied in a simplified economic system. *Econometrica* 2: 37–50.
1937. Indices and the simplified system. In *Report of third annual research conference on economics and statistics*, ed. Cowles Commission for Research in Economics. Chicago: University of Chicago Press.
1939. (With K. May.) Stability of limited competition and cooperation. In *Reports of a mathematical colloquium*, 2nd series. Notre Dame University.
1950. Mathematics for theoretical economists. *Econometrica* 18: 203–204.
1952. Note on the velocity of circulation of money. *Econometrica* 20: 1.
1954. Subjective values and value symbols in economics. In *From symbols and values, an initial study: Thirteenth symposium of the conference on science, philosophy, and religion.* New York: Harper & Row.

Evolutionary Economics

Ulrich Witt

Abstract

This article reviews the way of thinking about economic problems and the research agenda associated with the evolutionary approach to economics. This approach generally focuses on the processes that transform the economy from within and on their consequences for firms and industries, production, trade, employment and growth. The article highlights the major contributions to evolutionary economics and explains its key concepts together with some of their implications.

Keywords

Behavioural theory of the firm; Bounded rationality; Capital accumulation; Competition and selection; Darwin, C.; Diffusion of technology; Endogeneity and exogeneity; Entrepreneurship; Evolutionary economics; Evolutionary game theory; Georgescu-Roegen, N.; Growth; Industrial clusters; Innovation; Institutionalism; Institutions; Knowledge; Learning; Life

cycle; National innovation systems; Novelty; Natural selection; Path dependence; Preferences; Production theory; Productivity growth; Replicator equation; Research and development; Routines; Rules of conduct; Schumpeter, J.; Selection; Structural change; Technical progress; Technology; Veblen, T

JEL Classifications
A12; B25; B52; D01; D21; L10; O31; O33; O40

Evolutionary economics focuses on the processes that transform the economy from within and investigates their implications for firms and industries, production, trade, employment and growth.

These processes emerge from the activities of agents with bounded rationality who learn from their own experience and that of others and who are capable of innovating. The diversity of individual capabilities, learning efforts, and innovative activities results in growing, distributed knowledge in the economy that supports the variety of coexisting technologies, institutions, and commercial enterprises. The variety drives competition and facilitates the discovery of better ways of doing things. The question in evolutionary economics is therefore not how, under varying conditions, economic resources are optimally allocated in equilibrium given the state of individual preferences, technology and institutional conditions. The questions are instead why and how knowledge, preferences, technology, and institutions change in the historical process, and what impact these changes have on the state of the economy at any point in time.

Posing the questions this way has consequences for the way theorizing is done in evolutionary economics. First, preferences, technology and institutions become objects of analysis rather than being treated as exogenously given. Second, following from the very notion that evolution is a process of self-transformation, the causes of economic change are in part considered to be endogenous, and not exclusively exogenous shocks. More specifically, these causes are identified with the motivation and capacity of economic agents to learn and to innovate. Third, the evolutionary process in the economy is assumed to follow regular patterns on which explanatory hypotheses can be based, rather than forming an erratic sequence of singular historical events.

These three meta-premises are widely shared in evolutionary economics. However, the details of the argument, methods, and even the specification of the attribute 'evolutionary' vary, corresponding to the different theoretical traditions in which evolutionary economics is rooted. The concept of evolution has a long history in economics and social philosophy. This antedates – and, to a certain extent, has influenced – Darwin's theory of the origin of species by means of natural selection. Where the concept of evolution originally stood for a process of betterment (of human society), the Darwinian revolution in the sciences purged these progressive, teleological connotations. Today, evolutionary thought usually defines itself in relation to the Darwinian theory of evolution, the contributions to evolutionary economics not excepted. Some authors consider Darwinian theory to be the master theory. Others borrow from it at a heuristic level for their analogy-driven theorizing in economics. Yet others explicitly dissociate themselves from Darwinian thought.

Schumpeter and the Neo-schumpeterian Synthesis

Schumpeter avoided the term 'evolution'. He considered it a Darwinian concept and denied such concepts any economic relevance. However, in his theory of capitalist development, Schumpeter (1934) clearly subscribes to the three meta-premises above. The restructuring of the economy is explained as emerging endogenously from ever new waves of major innovations implemented by pioneering entrepreneurs with unique capabilities and motivation. Technology and the institutions of capitalism are endogenized. The transformation process of the economy is assumed to be governed by regular patterns, that is, cycles of investment and growth – booms and depressions – triggered by the innovations that occur 'in waves'

and diffuse throughout the economy in competitive imitation processes.

In Schumpeter (1942, p. 83) innovations that 'incessantly revolutionize the economic structure from within' remain central, but the innovating agents change. Previously viewed as achievements of unique promoter-entrepreneurs, innovations now appear as the routine output of trained specialists in large corporations. Correspondingly, the driving force of capitalist development is identified in the risky R&D investments of the large trusts – undertaken only if they expect proper returns to be earned. To protect these returns from being competed away immediately, the large, innovative corporations tend to engage in monopolistic practices. Such practices are incompatible with the ideal of perfect competition, but without them there would be significantly fewer R&D investments and innovations. Moreover, Schumpeter (1942, Ch. 8) claims that monopolistic practices work for only a limited time before innovations are eventually imitated or invalidated by rival innovations. Despite temporary monopolistic practices, competition by innovation thus boosts economic growth and raises prosperity more than fiercer price competition could ever do. This notion of 'Schumpeterian competition' induced a long debate about the relationships between firm size, market structure and innovativeness in which, however, the broader concept of endogenous economic change was lost from sight.

Endogenous change returns to centre stage in Nelson and Winter's (1982) neo-Schumpeterian restatement of evolutionary economics that blends Schumpeter's ideas with Darwinian concepts on the one hand and elements of the behavioural theory of the firm on the other. Schumpeter (1942) had not been specific about the innovative operations of the large corporations. To fill the gap, Nelson and Winter assume that, because of bounded rationality, firms operate on the basis of organizational routines. Different firms develop different routines for producing, investing, price setting, using profits, searching for innovations, and so forth, resulting in a diversity of competitive behaviours in the industry. By analogy with the principle of natural selection, Nelson and Winter argue that this diversity tends to be eroded whenever competing routines lead to differences in the firms' market performance and profitability. The better the firms perform, the more likely they are to grow, and the less reason they have to change their routines. The opposite holds for poorly performing firms. Much as differential reproductive success raises the share of better adapted genes in the gene pool of a population, differential firm growth thus raises the relative frequency of the better adapted routines in the 'routine pool' of the entire industry.

Instead of being a matter of optimal, deliberate substitution between given alternatives, in this view, the firms' competitive adaptations to changing market conditions are forced on them by selection processes operating on their routines. However, in a Schumpeterian spirit, Nelson and Winter also account for innovative moves – a breaking away from old routines – in an industry's response to changing market conditions. New ways of doing things, for example in responding to rising input prices, are established by search processes which are themselves guided by higher-level routines. Modelled as random draws from a distribution of productivity increments, innovations raise the average performance of the industry and regenerate the diversity of firm behaviours for selection to operate on. Some of the firms are driven out of the market, while the surviving ones tend to grow. Under innovation competition, technology and industry structure thus co-evolve and feed a non-equilibrating economic growth process. Regarding the debate on Schumpeterian competition, Nelson and Winter's analysis suggests a reversal of cause and effect: a high degree of concentration within an industry (an indicator of monopolistic power) may evolve as a consequence of, rather than being a prerequisite for, a high rate of innovativeness in the industry.

Selection Principles and Processes

Analogies between natural selection and market competition are not new. Better-adapted variants of firm behaviour have often been argued to prevail in an industry just as better-adapted variants

tend to prevail under natural selection pressure in the population of a species (an argument that has sometimes been misunderstood as vindicating profit-maximizing behaviour). The logic of the argument can be rendered more precise (Metcalfe 1994). Consider an industry with firms $i = 1,\ldots, n$ producing a homogeneous output with unit cost $c_i = \text{const}$. Assume that the firms use different organizational routines which result in a non-degenerate unit cost distribution. Let $s_i(t)$ denote the market share of firm i at time t measured by output. In a competitive market in which trade takes place at a uniform price $p(t)$,

$$p(t) = c(t) = \sum_i s_i(t) \cdot c_i, \qquad (1)$$

with $c(t)$ as the average level of unit cost in the industry. By Eq. 1, the average profit in the industry is zero. For at least one firm i, however, individual profit $\pi_i = p(t) - c_i > 0$ unless the entire market is served by the firm with the lowest level of unit cost.

Let the firm's growth be expressed in terms of the rate of change of its market share $(ds_i(t)/dt)/s_i(t)$ that is assumed to be a monotonic function ϕ of the firm's profit. With (Eq. 1) inserted into the individual profit equation, the rate of change of the firm's market share can therefore be written as

$$\frac{ds_i(t)/dt}{s_i(t)} = \phi(c(t) - c_i) = \phi(\pi_i(t) - \pi(t)). \quad (2)$$

Hence, performance differences across firms and their routines translate into corresponding differential growth rates of the firms.

The 'replicator' Eq. 2 corresponds to what is called 'Fisher's principle' in population genetics (Hofbauer and Sigmund 1988, Ch. 3). Let the fitness of an organism carrying a certain genetic trait be a constant. If it exceeds the average fitness in a population, the relative share of that trait in the population increases and vice versa. Consequently, natural selection raises average fitness over time to the level of the highest individual fitness. The change of the mean population fitness is proportional to the variance of the individual fitness. Analogously, with $c(t)$ as the measure for 'population fitness' in Eq. 2, $dc(t)/dt = f(\text{Var}(c_i)) \leq 0$.

If individual fitness is not constant, Fisher's principle no longer applies.

Suppose individual unit costs decrease with the firms' output, for example because of scale economies. The replicator equation can then have several fixed points representing multiple selection equilibria associated with a different average cost level (Metcalfe 1994). Which of the multiple equilibria the process converges to – and, consequently, whether the *ex ante* most profitable cost practice is eventually selected – depends on the initial conditions. Selection does not necessarily drive fitness or, for that matter, profits to the largest maximum. (Replicator equations with multiple equilibria can also result if the individual fitness terms depend on the population shares of their carriers. Such a frequency dependency is characteristic of models in evolutionary game theory; see Hofbauer and Sigmund 1988, Ch. 16).

To influence the underlying distribution of traits or behaviours, selection requires sufficiently inert conditions. In economic transformation processes this condition is often systematically violated. For example, firms facing a declining market share and/or profitability have strong incentives to modify their operations, that is, to replace inferior routines and/or to search for innovations. In general, with innovations playing a central role – as in Schumpeterian capitalist development – the volatility of the firms' environment increases and makes inertia rather unlikely. Industry dynamics are then more likely to be shaped by the generation and diffusion of innovations following their own time patterns rather than by selection processes. While in the case of selection processes theorizing focuses at the population level ('population thinking'), the explanation of the generation and diffusion of innovations can benefit from reconstructing motives and capabilities at the individual level.

Emergence and Diffusion of Innovations

Important as innovations are for economic transformation processes, the possibilities for

analysing how they emerge are limited because the underlying cognitive processes are basically unknown. What can nonetheless be analysed is why and when agents are motivated to search for innovations, provided their motivation is not made contingent on the – as yet unknown – outcome of the search (as in models of optimal choices between known alternatives that are therefore not applicable here). Often search motivation is triggered by a state of dissatisfaction or deprivation that the agents want to overcome by actions still to be found. Among the causes may be unsatisfied curiosity, a motivation to achieve something (Schumpeter 1934), or an agent's aspiration level that is temporarily not satisfied (Nelson and Winter 1982, Ch. 9). Where individual motivations like these occur in an uncorrelated way, they induce a base rate of innovative activity in the economy. If, in contrast, search motivation arises in a correlated way, for example in an economic crisis or when an industry is exposed to major innovations, the rate of innovative activities can rise far above the base rate. This is the case, for example, when firms need to innovate or be fast imitators with sufficient absorptive capacity in order to survive and therefore routinely engage in R&D.

Once an innovation is created or discovered by an agent, its implications can be grasped. Suppose, after assessing its benefits and costs, an agent implements an innovation. The implementation can usually be observed by competitors and/or other potential users. Since, in the absence of independent, own experience, people often draw conclusions from observing what others do, some observers may thus infer that the innovation is profitable and may start imitating it. Other observers may draw this conclusion only after a number of competitors and/or potential users have also signalled that they expect to benefit from adopting the innovation. Observational learning of this kind implies a dependency of the individual imitation or adoption behaviour – and, hence, the diffusion of the innovation – on the relative frequency of adopters.

The logic of this dependency can be captured by a function $q(t) = g(F(t))$, depicting the probability $q(t)$ that an agent who decides in t will adopt the innovation against the relative frequency of adopters $F(t)$ at time t. For $q(t) > F(t)$ the expected relative share of adopters grows with each additional decision and vice versa for $q(t) < F(t)$. The diffusion dynamics

$$\frac{dF(t)}{dt} = q(t) - F(t) \tag{3}$$

therefore hinge on the shape of the function g. For the quadratic function $q(t) = aF(t) - aF(t)^2$, $a > 1$, for instance, $F(t)$ converges to a fixed point F^a, $0 < F^a \leq 1$, that depends on the size of a. (By integration of Eq. 3 the diffusion path can in this case be shown to follow the well-known S-shaped logistic trend).

For the cubic function $q(t) = 3F(t)^2 - 2F(t)^3$, to take that example, the condition $q(t) = F(t)$ is satisfied if F equals 0, $\frac{1}{2}$, or 1. Inserting the cubic function into Eq. 3, F = 0 and F = 1 can be shown to represent stable fixed points of Eq. 3 while $F^* = \frac{1}{2}$ represents an unstable fixed point. This implies that for $F(t) < F^*$ the probability of adopting the innovation is too small to induce a spontaneous diffusion process. If $F(t)$ were for some reason to exceed F^* – representing a 'critical mass' of adopters – the innovation would however spread. The reason could be fluctuations of $F(t)$ that randomly cumulate, but are not represented in this simple deterministic model. (This explanation also plays a role in evolutionary game theory where the question is, for example, whether a new convention can emerge in a coordination game; see Young 1993). Another reason could be that somebody organizes a collective action by which the critical mass of agents is made to believe that more than the share F^* of agents will adopt the innovation.

With major technological innovations, competing variants or designs that serve the same user needs are often spawned simultaneously. The diffusion processes of the competing variants are interdependent if, for each of the variants, the users' utility varies with the number of adopters. Such 'economies to adoption' of alternative variants have been diagnosed, for example, for electric current transmission, video recorder systems, or the layout of typewriter keyboards.

The underlying pattern is again a frequency-dependency effect that can be analysed as before, if only two rival variants are assumed and the decision of agents who adopt neither of these is neglected.

Let $q(t)$ denote the probability of adopting the first variant and $F(t)$ its share of adopters at time t. Suppose both variants become available simultaneously and offer the same inherent benefits. For the first variant the development is captured by the cubic function above, interpreted as the mean process of a stochastic adoption process. With an identical number of initial adopters, $F(0) = F^* = \frac{1}{2}$ and $q(t) = \frac{1}{2}$. Once $F(t) \neq \frac{1}{2}$ for $t > 0$, economies to adoption raise the individual adoption probability of one of the variants over that of the other. As a consequence, the realization of the stochastic diffusion process initially fluctuates around F^*. Over time, however, small historical events and cumulative random fluctuations drive the process in the direction of either $F = 0$ (first variant disappearing) or $F = 1$ (second variant disappearing). In competitive diffusion processes of this kind, the prevailing state of the technology is thus 'path-dependent', and the process can be 'locked in' to the one variant if it is assumed, in addition, that over time the number of adopters grows beyond all bounds (Arthur 1994, Ch. 3). This means that, for $t \to \infty$, the likelihood of passing F^* by cumulative random fluctuations goes to zero.

The Evolution of Industries and the Institutions Backing Innovativeness

The substitution processes that the diffusion of new products and techniques induces shake up the established production structures. Factor owners and producers are forced to make adjustments – often painful ones that depreciate earlier investments and acquired competencies. While such 'pecuniary externalities' are inevitable concomitants of innovations, the longer-run consequence of innovativeness is – as Schumpeter (1942) had postulated – a rising standard of living of the masses. As a result of innovativeness, labour productivity and per capita income increase. New products and services absorb the growing consumption expenditures where established markets tend to be satiated. New employment opportunities emerge in new industries. To understand the working of the innovative transformation process and its policy implications, it is often useful to reconstruct the historical record of the evolution of entire industries (Malerba et al. 1999). Many of them, like the auto industry or the computer industry, grow out of a few major innovations for which new markets can be established or existing ones can substantially be expanded. Industries continue to grow over time under the pressure of imitative competition, often following a path of technical improvements that evolves within a 'technological paradigm' (cf. Dosi 1988).

Such regular patterns of change at the industry level can for many, though not all, industries be characterized in a stylized way by a life-cycle metaphor (Klepper 1997). Soon after their markets have been established by early innovators, the industries experience heavy entry and exit activities by competitors who partly imitate and partly add new varieties. While the market is expanding, a drastic shake-out in the number of firms occurs so that eventually a few large firms dominate the industry, and diversity in products and processes is reduced. In the beginning, product innovations are a main source of competitive advantages. Over time, however, the importance of process innovations increases. They raise productivity, drive down unit costs, and tend to intensify price competition. One cause of these patterns of industry evolution seems to be increasing returns to process innovations. These favour first movers that have been able to attain a sufficient size to spread development costs over larger output bases. With fiercer price competition, the firms with higher unit costs tend to be driven out of the industry, as in the selection model discussed above. Market concentration rises. With fewer innovations at that stage in the industry, its growth slows down, if the industry is not stagnating or declining.

Industry evolution is often connected with spatial effects. Innovative production techniques and

new products often grow out of initiatives, competencies, endowments, and institutional settings in particular locations (Antonelli 2001). If such complementary and interdependent local innovative activities gain momentum and trigger a self-augmenting process of firm growth and firm founding activities in close spatial proximity, an 'industrial cluster' can emerge. During early phases of the industry life cycle, a substantial share of the corresponding national or international industrial innovative activity may even be concentrated in such locations, Silicon Valley being the paradigmatic case. In such regions, income and employment are boosted. For policymaking the question therefore arises under what conditions innovative industrial clusters emerge and how and when their emergence can be fostered (Brenner 2004).

The early growth of innovative industries creates new employment opportunities. At later stages of the industry life cycle, when price competition and substitution pressure from innovative industries force the industry to raise labour productivity to reduce costs, employment is usually gradually lost. (For this reason, an industrial cluster that dominates a region can, in later stages of the industry life cycle, become a drag on local employment and prosperity). At the macroeconomic level, the stages reached in the life cycles of the industries interact in a complex way with productivity and income growth rates, and with the overall changes in employment (Metcalfe et al. 2006). Although these interactions have not yet been fully explored, it seems clear that at least two conditions must be met to maintain a high level of aggregate employment. First, innovative industries with new employment opportunities must emerge at the right times to compensate for the labour-saving technical progress. Second, the workforce must be able to adjust to the qualification requirements of the innovative industries and technologies. Since there is no self-regulating mechanism fulfilling the first condition, and because of delays and frictions in satisfying the second condition, the evolution of the industries is not necessarily a smooth transformation process. Aggregate employment and domestic income can vary substantially with the pace at which innovative industries emerge and expand. However, high levels of education and training are likely to raise innovativeness and the qualifications of the workforce. Ensuring this with an adequate institutional infrastructure – a productive national system of innovations – is an important policy option in supporting and smoothing the transformation process. This is even more true from a global perspective. A country's growth potential and its competitive advantage in trade hinge on when the country gains access to newly emerging technological opportunities and where in the innovative industries' life cycle it enters the market. History shows that differences between countries in this respect correspond to differences in their national innovation systems (Fagerberg 2002).

Darwinian Perspectives on Economic Evolution

The neo-Schumpeterian approach considers the concept of selection as constitutive for evolutionary economics. Economic selection processes, operating on the diversity of individual behaviours, force adaptations on populations of agents who are prevented by their bounded rationality from deliberately adapting optimally. The import of the selection concept is not meant to extend Darwinism to the economic domain. Such an extension was, however, advocated by Veblen (1898) under the influence of the Darwinian revolution of his time. He coined the term 'evolutionary' economics for such an approach (Hodgson 2004). A Darwinian perspective on the economic domain can indeed help to clarify how evolutionary economics fits with the Darwinian world view now prevailing in the sciences and in this way offer new insights (Witt 2003).

In the economic domain, the bulk of change to be explained occurs within single generations. In contrast, the Darwinian theory of natural selection focuses on inter-generational change and is therefore relevant only for explaining the basis on which economic evolution rests. These are, first, the long-term constraints man-made economic evolution is subjected to and, second, the innate dispositions and adaptation mechanisms in

humans (shaped earlier in human history by natural selection) that define the basic behavioural repertoire. Veblen (1898) focused on habits, including habits of thought, which he assumed to emerge from hereditary traits and past experiences, given the traditions, conventions and material circumstances of the time. (Habits play the crucial role in Veblen's explanation of the 'cumulative causation' of institutions which, in turn, he regarded as the key to understanding the different forms of economic life and their genesis).

In a similar vein one may focus on human preferences that emerge from the interplay of inherited dispositions and innate conditioning learning mechanisms – both of these shared by all humans with the usual genetic variance. A prominent example of innate dispositions is the altruistic attitudes that play a prominent role in evolutionary game theory (Hofbauer and Sigmund 1988, Ch. 14). Other examples of innate dispositions can be found in certain forms of consumption. The genetically fixed learning mechanism accumulates the influence of a lifelong history of reinforcement and conditioning. It is responsible for the emerging variety of individual preferences and keeps them changing over time.

Following Hayek (1988, Ch. 1), innate behaviour can be conjectured to play a key role in the evolution of human institutions. They emerge, he argues, through social learning of 'rules of conduct' that starts from primitive, genetically fixed, forms of social behaviour and add on new elements by trial and error. Over their history, different groups or whole societies thus build up a diversity of rules that regulate their interactions. The group members' innovativeness is channelled into economic activities provided institutional regulations do not discourage this or fail to protect the capital accumulation that is necessary to realize innovations. Those groups that succeed in developing and passing on rules able to better meet these conditions can therefore be expected to grow and prosper in terms of population size and per capita income. Their differential success may enable such groups to conquer and/or absorb less well-equipped, competing groups and thus propagate better adapted institutions.

Economic evolution is, of course, also shaped in an essential way by human intelligence. By cognitive learning, problem solving and inventiveness, knowledge about institutions, opportunities and technologies is created (Mokyr 2002). In the longer run, the enabling effects of cumulative knowledge generation emerging over time matter more than the effects of economizing on scarce resources at each point in time. From a Darwinian perspective the most significant tendency in the use of cumulative knowledge is the manipulation of natural constraints to better accord them with human preferences. This has enlarged the niche for the human species and has improved living conditions for an ever-increasing number of its members. At the same time, however, knowledge accumulation has contributed to dramatically increasing the human share in the use of natural resources. According to Georgescu-Roegen's (1971) evolutionary approach to production theory, this way of solving problems implies a risky long-term impact on nature, the ultimate basis of the human economy. To account for these risks further innovative efforts that transform the economy from within seem indispensable.

See Also

- ▶ Analogy and Metaphor
- ▶ Competition and Selection
- ▶ Deterministic Evolutionary Dynamics
- ▶ Diffusion of Technology
- ▶ Learning and Evolution in Games: An Overview
- ▶ Path Dependence
- ▶ Schumpeter, Joseph Alois (1883–1950)
- ▶ Schumpeterian Growth and Growth Policy Design
- ▶ Structural Change
- ▶ Veblen, Thorstein Bunde (1857–1929)

Bibliography

Antonelli, C. 2001. *The microeconomics of technological systems*. Oxford: Oxford University Press.

Arthur, W. 1994. *Increasing returns and path dependence in the economy*. Ann Arbor: Michigan University Press.

Brenner, T. 2004. *Local industrial clusters – existence, emergence and evolution*. London: Routledge.
Dosi, G. 1988. Sources, procedures, and microeconomic effects of innovation. *Journal of Economic Literature* 26: 1120–1171.
Fagerberg, J. 2002. *Technology, growth and competitiveness*. Cheltenham: Edward Elgar.
Georgescu-Roegen, N. 1971. *The entropy law and the economic process*. Cambridge, MA: Harvard University Press.
Hayek, F. 1988. *The fatal conceit*. London: Routledge.
Hodgson, G. 2004. *The evolution of institutional economics*. London: Routledge.
Hofbauer, J., and K. Sigmund. 1988. *The theory of evolution and dynamical systems*. Cambridge: Cambridge University Press.
Klepper, S. 1997. Industry life cycles. *Industrial and Corporate Change* 6: 145–181.
Malerba, F., R. Nelson, L. Orsenigo, and S. Winter. 1999. 'History-friendly' models of industry evolution: The computer industry. *Industrial and Corporate Change* 8: 3–40.
Metcalfe, J. 1994. Competition, Fisher's principle and increasing returns in the selection process. *Journal of Evolutionary Economics* 4: 327–346.
Metcalfe, S., J. Foster, and R. Ramlogan. 2006. Adaptive economic growth. *Cambridge Journal of Economics* 30: 7–32.
Mokyr, J. 2002. *The gifts of athena*. Princeton: Princeton University Press.
Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge: Harvard University Press.
Schumpeter, J. 1934. *The theory of economic development*. Cambridge: Harvard University Press.
Schumpeter, J. 1942. *Capitalism, socialism and democracy*. New York: Harper.
Veblen, T. 1898. Why is economics not an evolutionary science? *Quarterly Journal of Economics* 12: 373–397.
Witt, U. 2003. *The evolving economy*. Cheltenham: Edward Elgar.
Young, P. 1993. The evolution of conventions. *Econometrica* 61: 57–84.

Ex Ante and Ex Post

Otto Steiger

Keywords

Aggregate demand and supply; Ex ante and ex post; Expectations; Lindahl, E. R.; Myrdal, G.; Ohlin, B. G.; Stockholm School; Temporary equilibrium

JEL Classifications

B4

The concepts of *ex ante* and *ex post* are the most popular terminological innovations developed by the famous so-called Stockholm School in the 1930s. The terminology was introduced into macroeconomic theory, especially with regard to the savings-investment relation by Gunnar Myrdal (1933, 1939) and clarified and incorporated into sequence or period analysis by Erik Lindahl (1934, 1939b), whose conceptual system of 'prospective' and 'retrospective' values achieved 'world-citizenship' as a method for drawing up national budgets (Hansen 1951, p. 27). The popularization of the method of *ex ante* and *ex post* is due to Ohlin's seminal articles on the Stockholm School (1937) which made it 'generally accepted over the whole world with a rapidity unusual to economics' (Palander 1941, p. 34).

The significance of the distinction between *ex ante* and *ex post* 'as one of the most transforming insights that theoretical economics has had' (Shackle 1972, p. 440) does not follow so much, as often stressed in the literature, from the simple fact that there exist always two alternative definitions of flow-related economic magnitudes like income, production, and so on, depending on whether they are looked at 'from before' or 'from after'. The central idea of the necessity to distinguish between *ex ante* and *ex post* stems rather from the recognition of the fundamental difference, originally expressed by Frank Knight (1921, pp. 35 f.) and definitely formulated by Myrdal (1939, pp. 59 f.), between 'foreseen' and 'unforeseen' changes where only the latter result in 'gains and losses' which, as shown by Lindahl (1939b, pp. 103 f.), have to be 'windfalls'. Therefore, in the analysis of expectations under uncertainty time has to be included in an essential way by two alternative methods of calculation of economic variables: (i) an *ex ante* computation or business calculation which refers to a point of time at the beginning of a period and (ii) an *ex post* computation or bookkeeping referring to the development in time at the end of the period (Myrdal 1939, pp. 45–7). As a consequence,

economic analysis can be divided into (i) an *ex ante* analysis explaining how expectations determine an economic magnitude and (ii) an *ex ante*/*ex post* analysis explaining the possible divergence between the expected and realized value of this variable.

The emergence of the concepts of *ex ante* and *ex post* can be dated to Lindahl's and Myrdal's early writings in the 1920s. In his first treatise in macroeconomic theory (Lindahl 1924, ch. 3) Lindahl stressed the time factor for economic analysis and used the notion of 'subjective calculations of the future' as well as the term *ex post* when he discussed 'a negative investment recognized only ex post' (p. 33). A more coherent analysis of these concepts was given in Myrdal's dissertation on expectations and price changes (Myrdal 1927, pp. 67 f.) where he showed, emphasizing Knight's idea of the difference between certain and uncertain changes, how divergences between incomes and costs of an investment calculated '*before*' will be balanced by gains and losses calculated '*after*'.

The first application of these ideas to macroeconomic problems was made by Lindahl in *Penningpolitikens medel* (Lindahl 1930; cf. 1939a). However, the dynamic method of temporary equilibrium used in this treatise, that is, 'an analysis dividing time into a number of short equilibrium periods during which no changes occur', led to a 'theoretically inadmissible mixture of the *ex ante* and *ex post* analysis' (Myrdal 1939, p. 122). In case of the same but wrong expectations, for example, the equality between savings and investment *ex ante* is not a guarantee for temporary equilibrium (Palander 1941, p. 44; see also Siven 2006, pp. 684–5). As shown by Myrdal in the original Swedish version of *Monetary Equilibrium* (Myrdal 1932, pp. 228–30), this mixture was especially obvious in Lindahl's discussion of the relation between investment and saving, where he could not demonstrate in a satisfactory way how an initial discrepancy due to a shift in the rate of interest will always be balanced by changes in the distribution of income between borrowers and lenders. If, however, Lindahl would give up his method of temporary equilibrium, that is, allow for disequilibrium during a period, his analysis could be interpreted in an *ex ante*/*ex post* framework.

It was exactly this disequilibrium analysis which enabled Myrdal to clarify the relation between investment and saving in his three different versions of *Monetary Equilibrium*, where he introduced the notions of *ex ante* and *ex post* first in the German edition (Myrdal 1933, § 29). In his discussion of these concepts Myrdal (1939, pp. 59–62, 116–25; cf. 1933, §§ 32, 55–6) allowed for a discrepancy between investment and saving *ex ante* based on 'anticipatory calculations' at a point of time demarcating the beginning of a period, while at its end their values were constructed by 'a subsequent "bookkeeping" in such a way that there is *always* an *ex post* balance 'regardless of how short the period'. Therefore, it is not 'this meaningless balance' which is of interest to economic analysis but '*the very changes during the period which are required to bring about this ex post balance*'. Myrdal assumed that these balancing factors arise out of 'unanticipated changes' in 'revenues and costs', that is, in incomes, during the period for which they can be calculated only *ex post*: gains and losses. The reason why *ex ante* and *ex post* values may differ is that expectations formulated in the beginning of a period are *ex ante* values of prices and quantities that may not be realized because expectations may be disappointed during the period (Siven 2006, p. 681). As later shown by Lindahl (1939b, pp. 103 f.; cf. 1958), these values have to be windfalls and must not be confused, as sometimes in Myrdal's analysis (Myrdal 1939, p. 65), with entrepreneurial gains or losses, which are already included in the *ex ante* values and which, therefore, cannot serve as balancing factors.

Although Myrdal always spoke of 'income changes' as the balancing factor *ex post* of discrepancies *ex ante* between investment and saving, his examples implied almost exclusively '*price changes*' (Myrdal 1939, p. 60; cf. Palander 1941, pp. 42f.; Hansson 1982, p. 149). It was left to Lindahl (1934, 1939b) to demonstrate how the *ex post* equality was achieved in a disequilibrium process via a change in quantities.

Lindahl presented his solution in an aggregate demand and supply framework, where demand

was identified with the purchase plans of consumers and producers and supply with the expectations of the sellers of production and consumption goods. In his analysis he made the fundamental assumption that the purchase plans as well as the supply prices of the sellers at the beginning of a period 'have been actually realized during the period' (Lindahl 1934, p. 207, cf. Lindahl 1939b, p. 92). Under this assumption a possible deviation between expected and realized sales, which Lindahl took as given if the future is not foreseen with certainty, must be considered as a result of a difference between investment and saving *ex ante* which in turn will cause a divergence between expected and realized total real income. These changes in income represent gains or losses to the producers which in a form of 'unintentional', not 'forced', saving or dissaving equalize investment and saving *ex post.*

The purpose of Lindahl's rigid assumption that purchase plans are always fulfilled, which made it impossible to apply his analysis to the conditions of full employment (Hansen 1951, pp. 29–32), was to demonstrate that, once prices are given, the *actions* of the economic subjects during the period 'can be directly deduced from the plans at the beginning of the period' (Lindahl 1939b, p. 92). With this demonstration Lindahl had taken the first step to a sequence analysis, that is, a single-period analysis where *ex ante* plans determine *ex post* results. The second step consisted of a continuation analysis where the *ex post* events of the current period lead to revisions of the *ex ante* plans for the consecutive period at the transition point between these periods, 'especially as regards the supply prices and the producers' and consumers' demand' (Lindahl 1934, p. 211). However, as Lindahl 'never succeeded in formulating 'laws of motion' for revisions of plans ... this promising branch of dynamic theory became abortive' (Hansen 1966, p. 3).

Of greater influence for economic analysis was Lindahl's second contribution to the development of the *ex ante*/*ex post* method, his discussion of the relations between 'prospective' and 'retrospective' values of micro- and macroeconomic variables (Lindahl 1939b) which contained 'the germ of many lines in later works' on the methodology of national accounting (Ohlsson 1953, p. 266). Although Lindahl's accounting structure was criticized as 'deficient' (Ohlsson) in the treatment of government accounting, it has been emphasized recently by Hicks (1985, p. 80) that Lindahl's system 'does have some continuing merits ... for the accounting of the public sector': 'In this field at least, it may still be contended, *ex ante ex post* remains respectable.'

For a long time it was argued that the exposition of the Keynesian system 'requires the language of *ex ante* and *ex post*' (Shackle 1972, p. 172; see also Patinkin 1976, pp. 139–40; Siven 2006, p. 700), with the emphasis placed on the possible divergences, due to uncertainty, between disappointed *ex ante* expectations and *ex post* results as one of the relevant factors in determining the level of employment. However, the posthumous publication of Keynes's 1937 lecture notes have shown that Keynes (1937a, p. 183; see also 1937b) emphasized that even under the assumption of an 'identity of *ex post* and *ex ante*', that is, with expectations always fulfilled but without having to assume for this case, as did Myrdal and Lindahl, the absence of uncertainty, 'the theory of effective demand is substantially the same' (p. 181). Moreover, as Keynes regarded the '*time* relationship' between the concepts of *ex ante* and *ex post* as 'incapable of being made precise' (p. 179), he rejected this method as an inadequate tool in handling the problems of uncertainty and time.

See Also

- Lindahl, Erik Robert (1891–1960)
- Myrdal, Gunnar (1898–1987)
- Stockholm School

Bibliography

Hansen, B. 1951. *A study in the theory of inflation*. London: Allen & Unwin.

Hansen, B. 1966. *Lectures in economic theory. Part I: General equilibrium theory*. Lund: Studentlitteratur.

Hansson, B. 1982. *The Stockholm school and the development of dynamic method*. London: Croom Helm.

Hicks, J. 1985. *Methods of dynamic economics*. Oxford: Clarendon Press.
Keynes, J.M. 1937a. Ex post and ex ante. Notes from the 1937 lectures. In *The general theory and after, Part II: Defence and development, vol. 14 of the collected writings of John Maynard Keynes*, ed. D. Moggridge. London: Macmillan.
Keynes, J.M. 1937b. The 'ex-ante' theory of the rate of interest. *Economic Journal* 47: 663–669.
Knight, F.H. 1921. *Risk, uncertainty, and profit*, 1971. Chicago: University of Chicago Press.
Lindahl, E. 1924. *Penningpolitikens mål och medel. Del I [The aims of means of monetary police. Part I]*. Lund: Gleerup; Malmö: Försäkringsaktiebolaget.
Lindahl, E. 1930. *Penningpolitikens medel [The means of monetary policy]*. Lund: Gleerup; Malmö: Försäkringsaktiebolaget; enlarged version of 1st edn, 1929; revised version trans. as Lindahl (1939a).
Lindahl, E. 1934. A note on the dynamic pricing problem. Mimeo, Gothenburg, 13; quoted from the corrected version published in Steiger (1971).
Lindahl, E. 1939a. The rate of interest and the price level. In Lindahl (1939c); revised version of Lindahl (1930).
Lindahl, E. 1939b. Algebraic discussion of the relations between some fundamental concepts. In Lindahl (1939c).
Lindahl, E. 1939c. *Studies in the theory of money and capital*. London: Allen & Unwin.
Lindahl, E. 1958. The concept of gains and losses. In *Festskrift til Frederik Zeuthen*. Copenhagen: Nationalkonomisk Forening.
Myrdal, G. 1927. *Prisbildningsproblemet och föränderligheten [The problem of price formation and changeability]*. Uppsala: Almqvist & Wiksell.
Myrdal, G. 1932. Om penningteoretisk jämvikt. En studie över den 'normala räntan' i Wicksells penninglära [On monetary equilibrium. A study on the 'normal rate of interest' in Wicksell's monetary thought]. Ekonomisk Tidskrift 33(5–6) (1931; printed 1932), 191–302; revised version trans. as Myrdal (1933).
Myrdal, G. 1933. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie*, ed. F.A. Hayek. Vienna: J. Springer; 1st revised version of Myrdal (1932); 2nd revised version trans. as Myrdal (1939).
Myrdal, G. 1939. *Monetary equilibrium.* London: Hodge; revised version of Myrdal (1933).
Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment I–II. *Economic Journal* 47 (53–69): 221–240.
Ohlsson, I. 1953. *On national accounting*. Stockholm: Konjunkturinstitutet.
Palander, T. 1941. Om 'Stockholmsskolans' begrepp och metoder. Metodologiska reflexioner kring Myrdals 'Monetary Equilibrium'. *Ekonomisk Tidskrift* 43(1): 88–143; quoted from and trans. as 'On the concepts and methods of the "Stockholm School": some methodological reflections on Myrdal's "Monetary Equilibrium". International Economic Papers No. 3 (1953), 5–57.
Patinkin, D. 1976. *Keynes's monetary thought: A study of its development*. Durham: Durham University Press.
Shackle, G.L.S. 1972. *Epistemics & economics. A critique of economic doctrines*. Cambridge: Cambridge University Press.
Siven, C.-H. 2006. Monetary equilibrium. *History of Political Economy* 38: 668–709.
Steiger, O. 1971. *Studien zur Entstehung der Neuen Wirtschaftslehre in Schweden. Eine Anti-Kritik*. Berlin: Duncker & Humblot.

Examples

James Bonar

Examples in economics, as elsewhere, are simply cases, real or fictitious, or partly both, supposed to embody a general principle. They may be classified as follows: (1) *Real but general*, as Ricardo's hunters (*Principles*), and Adam Smith's bricklayers, carpenters, and men of letters (*Wealth of Nations*). The examples are taken from a known genus but not from known individuals. Where the genus is perfectly well known, no cavil is possible. Adam Smith's illustration of division of labour could hardly have been improved by a reference to a particular pin-making establishment in a specified place. But, in exposition, the more concrete the genus the more telling the example; e.g. 'blacksmith' seems nearer life than 'workman'. (2) *Real and particular*, as in Cairnes's illustration of the theory of international trade from the Australian gold discoveries. Adam Smith, where he does not use the real and general, uses the real and particular, and falls back on fiction only for his similes (as 'the highway', 'the waggonway through the air', the 'wings', and 'the pond and the buckets', *Wealth of Nations*, II, ii), or his metaphors ('wheel of circulation', 'channel of circulation'). Ricardo and his immediate followers have preferred, as a rule, (3) *Fictitious* examples. These may be illustrations of which the component elements are generically well known, even the favourite 'man on the desert

island', but the combining of the elements is the work of the writer, and is more or less arbitrary, as de Quincey's 'man with the musical box on Lake Superior', and Bastiat's 'plank and plane'. There is also a risk that the construction of the example may involve a begging of the question to be proved. 'Suppose that there are but two nations in the world living side by side, with a population of one million souls in each' (Barbour, *Bimetallism*). 'My object was to elucidate principles, and to do this I imagined strong cases that I might show the operation of those principles' (Ricardo, *Letters*). There is no necessary fallacy in this method of exposition any more than in illustrating the law of gravitation by the action of bodies *in vacuo*. Concrete cases must necessarily exemplify much more than one principle, and, even if they suggested a particular generalization, they may perhaps not clearly illustrate it without a fictitious simplification. The lawfulness of such a method of exposition or, it may be, of proof is discussed elsewhere.

Reprinted from Palgrave's Dictionary of Political Economy.

Excess Burden of Taxation

James R. Hines Jr.

Abstract

The excess burden of taxation is the efficiency cost, or deadweight loss, associated with taxation. Excess burden is commonly measured by the area of the associated Harberger triangle, though accurate measurement requires the use of compensated demand and supply schedules. The generation of empirical excess burden studies that followed Arnold Harberger's pioneering work in the 1960s measured the costs of tax distortions to labour supply, saving, capital allocation, and other economic decisions. More recent work estimates excess burdens based on the effects of taxation on more comprehensive measures of taxable income, reporting sizable excess burdens of existing taxes.

Keywords

Dupuit, A.-J.-E. J.; Excess burden of taxation; Harberger triangle; Jenkin, H. C. F.; Path dependence; Tax evasion; Taxable income

JEL Classifications

H3

The excess burden of taxation is the efficiency cost, or deadweight loss, associated with taxation.

The total economic burden of a tax includes both payments that taxpayers make to the government and any lost economic value from inefficient activities undertaken in reaction to taxes. Since direct tax burdens take the form of revenue that taxpayers remit to governments, the excess burden of taxation is the magnitude of the economic costs of accompanying economic distortions. For example, a tax on labour income typically discourages work by encouraging inefficient substitution of untaxed leisure for taxed paid work. At low tax rates this substitution entails only modest excess burdens, since, in the absence of other distortions, the welfare cost of substituting an untaxed for a taxed activity simply equals the tax rate, the difference between pre-tax and after-tax returns to the taxed activity. At high tax rates this difference is quite large, and as a result residents of economies with high tax rates may face substantial excess burdens of taxation. Indeed, it is entirely possible for the excess burden of a tax to exceed the revenue collected; a tax imposed at so high a rate that it eliminates the taxed activity clearly has this feature.

The excess burden of taxation is commonly measured by the area of the associated 'Harberger triangle' (Hines 1999). The base of the Harberger triangle is the amount by which economic behaviour changes as a result of price distortions introduced by the tax, and the height of the Harberger triangle is the magnitude of the tax burden per unit of economic activity.

The Many Excess Burdens

One of the difficulties that arise in evaluating the excess burden of taxation is that there is more than one possible measure of excess burden. This multiplicity does not imply that all measures are equally desirable or useful. For example, the use of uncompensated (Marshallian) demand and supply curves to construct Harberger triangles produces measures of the excess burden of taxation with a number of known problems. In the (realistic) case in which a government uses multiple taxes, a measure of total excess burden based on uncompensated demand and supply curves is path dependent, meaning that its value depends on the order in which the taxes are imagined to be imposed. As the order of the taxes is perfectly arbitrary, path dependence is troubling – most importantly because it reflects the imprecision of excess burden measures constructed in this way.

Path dependence is one consequence of this imprecision; another is that a tax system that produces a higher level of economic welfare might have a greater measured excess burden than an alternative that raises the same revenue. If excess burden is to be useful in the evaluation and formation of tax policies, it is necessary that the measure should correspond, at least approximately, to the economic cost of taxation – and assign greater excess burden to tax systems that are in fact more burdensome.

Path dependence and inaccurate welfare orderings need not arise if excess burden is measured by Hicksian consumers' surplus, based on schedules that hold utility, rather than income, constant as prices vary. Because actual tax policy changes typically do not hold utility constant, it is necessary to construct a measure based on a conceptual experiment that does. One intuitive experiment is to imagine that, as a tax is imposed, utility is held constant at its pre-tax level. Excess burden is then defined as the amount, in excess of tax revenue, that the government must compensate consumers to maintain initial utility in the face of a tax-induced price change. The amount of compensation, which corresponds to the Hicksian measure of the *compensating variation* of the price change, may be calculated in roughly the same way that Harberger triangles are commonly measured.

An alternative conceptual experiment is to begin with the tax already in place and then remove it, extracting from consumers in lump-sum fashion an amount that prevents them from changing their utility levels while the tax is removed. Because the initial tax is distortionary, it is necessary to extract more from consumers than the tax revenue, the difference representing the excess burden of the initial tax. This differs from the previous measure in corresponding to a Hicksian *equivalent variation* measure of excess burden. One virtue of an equivalent variation measure of excess burden, compared to the compensating variation measure, lies in the fact that, in comparing tax systems that raise equal revenue, the tax system with the lowest excess burden as measured by equivalent variation also produces the highest level of consumer welfare (Kay 1980).

Although these compensating variation and equivalent variation measures are the most intuitive, they are actually just examples drawn from a class of measures based on arbitrary levels of utility and arbitrary reference price vectors. As King (1983) and others note, the use of compensated supply and demand schedules together with fixed reference price vectors guarantees that resulting excess burden measures have desirable properties, though the interpretation of the resulting magnitudes depends on the choice of utility levels and price vectors. These measures then can be naturally generalized to include marginal excess burden, the change in excess burden arising from a given tax change, and to treat excess burden in settings in which costs of production vary with output levels (Auerbach and Hines 2002).

Empirical Measurement of Excess Burden

While the theory of excess burden measurement has a long and colourful history that dates back to the 19th century contributions of Jules Dupuit

(1844) and Fleeming Jenkin (1871–2), economists seldom measured actual excess burdens prior to the pioneering work of Arnold Harberger in the 1960s. In two influential papers published in 1964, Harberger (1964a, b) derived an approximation used to measure excess burden and (1964b) applied the method to estimate excess burdens of income taxes in the United States. Harberger shortly thereafter (1966) produced estimates of the excess burden of US capital taxes. A generation of empirical studies by other scholars followed the publication of Harberger's subsequent survey article (1971).

The empirical work that followed Harberger's efforts focused on the use of simple excess burden formulas to estimate the welfare impact of a wide array of tax- induced distortions, including those to labour supply (Browning 1975), saving (Feldstein 1978), corporate taxation (Shoven 1976), and the consumption of goods, such as housing and non-housing consumption items, that are taxed to differing degrees (King 1983). In addition, some attention was devoted to refining the approximations used in applying estimated behavioural parameters to calculate excess burdens. A variant of the excess burden formula used by Harberger, in which a form of uncompensated demand is used in place of compensated demand, approximates a compensated measure of welfare change. One question of interest to subsequent investigators is the practical difference between results obtained using Harberger-style approximations and those available from more exact measures. As Mohring (1971) and subsequent authors note, it is often the case that the same demand information necessary to calculate approximations can, if properly modified, be used to calculate Hicksian excess burden measures. The extent to which these two methods generate different answers is, of course, an empirical question. Rosen (1978) finds that measures of excess burden based on compensated and uncompensated demand and supply schedules track each other rather closely, but Hausman (1981) offers some examples in which they differ considerably.

A major practical difficulty in measuring the excess burden of a single tax, or of a system of taxes, is that excess burden is a function of interactions that are potentially very difficult to measure. For example, a tax on labour income is expected to affect hours worked, but may also affect the accumulation of human capital, the intensity with which people work, the timing of retirement, and the extent to which compensation takes tax-favoured (for example, pensions, health insurance, and workplace amenities) in place of tax-disfavoured (for example, wage) form. In order to estimate the excess burden of a labour income tax, it is in principle necessary to estimate the effect of the tax on these and other decision margins. Analogous complications are associated with estimating the excess burdens of most other taxes. In practice, it can be very difficult to obtain reliable estimates of the impact of taxation on just one of these variables.

It is in reaction to the complicated nature of the problem of separately estimating the effect of taxation on all of a taxpayer's decision margins that a number of recent studies estimate excess burdens based on the effects of taxation on reported taxable income. Taxable income incorporates not only any effects of taxation on work effort, but also tax avoidance of various forms, including deliberate hiding of income and legal avoidance such as making tax-deductible charitable contributions. Properly measured, excess burden, as calculated by the effect of taxation on taxable income, should accurately capture all the necessary interactions to evaluate the welfare consequences of taxation (Feldstein 1999).

Several empirical studies, including Feldstein (1995), Auten and Carroll (1999), and Goolsbee (2000), consider the responsiveness of taxable income to tax rates, relying on major US tax changes to provide variation in tax rates. The evidence indicates that taxable income is generally quite responsive to tax changes, particularly among the high-income population, thereby implying an excess burden of US taxes considerably greater than that produced by studies using estimated effects of taxation on work hours and saving. The estimates suggest excess burdens of taxation that might be as high as 75 per cent of tax revenue collected (Feldstein 2006), though there is still considerable uncertainty over its true magnitude.

See Also

- ▶ Neutral Taxation
- ▶ Optimal Taxation
- ▶ Pigouvian Taxes

Bibliography

Auerbach, A.J., and J.R. Hines Jr. 2002. Taxation and economic efficiency. In *Handbook of public economics*, 3rd ed, ed. A.J. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Auten, G., and R. Carroll. 1999. The effect of income taxes on household income. *Review of Economics and Statistics* 81: 681–693.

Browning, E.K. 1975. Labor supply distortions of social security. *Southern Economic Journal* 42: 243–252.

Dupuit, A.J.É.J. 1844. De la mesure de l'utilité des travaux publics. *Annales des ponts et chaussées,* 2nd ser., 8. Tr. R.H. Barback as 'On the measurement of the utility of public works'. *International Economic Papers* 2 (1952), 83–110. Reprinted in *Readings in welfare economics,* ed. K.J. Arrow and T. Scitovsky. Homewood: Richard D. Irwin, 1969.

Feldstein, M. 1978. The welfare cost of capital income taxation. *Journal of Political Economy* 86: S29–S51.

Feldstein, M. 1995. The effect of marginal tax rates on taxable income: A panel study of the 1986 tax reform act. *Journal of Political Economy* 103: 551–572.

Feldstein, M. 1999. Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics* 81: 674–680.

Feldstein, M. 2006. *The effects of taxes on efficiency and growth*, Working paper no. 12201. Cambridge, MA: NBER.

Goolsbee, A. 2000. What happens when you tax the rich? Evidence from executive compensation. *Journal of Political Economy* 108: 352–378.

Harberger, A.C. 1964a. The measurement of waste. *American Economic Review* 54: 58–76.

Harberger, A.C. 1964b. Taxation, resource allocation, and welfare. In *The Role of direct and indirect taxes in the Federal revenue system*, ed. J.F. Due. Princeton: Princeton University Press.

Harberger, A.C. 1966. Efficiency effects of taxes on income from capital. In *Effects of corporation income tax*, ed. M. Krzyzaniak. Detroit: Wayne State University Press.

Harberger, A.C. 1971. Three basic postulates for applied welfare economics: An interpretive essay. *Journal of Economic Literature* 9: 785–797.

Hausman, J.A. 1981. Exact consumer's surplus and deadweight loss. *American Economic Review* 71: 662–676.

Hines Jr., J.R. 1999. Three sides of Harberger triangles. *Journal of Economic Perspectives* 13: 167–188.

Jenkin, H.C.F. 1871–2. On the principles which regulate the incidence of taxes. *Proceedings of the Royal Society of Edinburgh* 7: 618–631. Reprinted in *Papers, literary, scientific, &c. by the late Fleeming Jenkin*, ed. S. Colvin and J. A. Ewing, vol. 2. London: Longmans Green, 1887.

Kay, J.A. 1980. The deadweight loss from a tax system. *Journal of Public Economics* 13: 111–120.

King, M.A. 1983. Welfare analysis of tax reforms using household data. *Journal of Public Economics* 21: 183–214.

Mohring, H. 1971. Alternative welfare gain and loss measures. *Western Economic Journal* 9: 349–368.

Rosen, H.S. 1978. The measurement of excess burden with explicit utility functions. *Journal of Political Economy* 86: S121–S135.

Shoven, J.B. 1976. The incidence and efficiency effects of taxes on income from capital. *Journal of Political Economy* 84: 1261–1283.

Excess Demand and Supply

Michael Allingham

In general equilibrium theory the economy may be represented by the function which specifies the aggregate excess demands (positive) or excess supplies (negative) which are expressed at all possible price systems.

While the concept of excess demand is implicit in Walras's (1874) framework it is first introduced explicitly in Hicks's (1946) treatment of the general equilibrium system. In the present discussion we first examine how the excess demand function is obtained from the underlying parameters of the economy, that is the preferences and endowments of the various agents in the economy. We then discuss why the concept is useful, and what restrictions economic theory imposes on the excess demand function, and, equally importantly, what it does not.

An economy with n commodities consists of a number of agents, each with given preferences for and endowments of these commodities. An agent's preferences are represented by a complete preordering P on the commodity space R^n_+.

The preordering has the following properties:

Continuity: the set of all x in R^n_+ such that xPy and the set such that yPx are both closed for all y.

Monotonicity: if $x \geq y$ but xy then xPy but not yPx.

Convexity: if xPy and z is a proper linear combination of x and y then zPy but not yPz. The agent's endowment is represented simply by a point e in $S = R^n_+ - 0$.

Given an agent's preferences P and endowment e, and given any price system p in S, there is one and only one $x(P)$ in $R^n{}_+$ such that $x(p)Py$ for all y such that $p \cdot y \geq p \cdot e$; existence of this element can be shown using Weierstrass's theorem, while uniqueness follows immediately from convexity (Debreu 1959). The difference between this $x(p)$ and his endowment that is $x(p) - e$, is his excess demand at the price system p, denoted $g(p)$.

One property which follows immediately from the definition of excess demand is that $g(p)$ is homogeneous (of degree zero), that is $g(tp) = g(p)$ for all positive t. A second property which follows immediately from the definition and monotonicity is that $p \cdot g(p) = 0$ for all p.

A third important property which is less immediate is that g is continuous provided that the value of the agent's endowment, $p \cdot e$, is positive (Debreu 1959). The reason why we need this proviso can be seen by an example. Let $n = 2$ and $e = (1,0)$, and let p tend to (0,1); for all positive p, $g_1(p)$ is non-positive but in the limit when $p = (0, 1)$ g_1(p) is infinite because of monotonicity.

A further property of excess demand is the revealed preference property. Consider two distinct prices p and q, with the corresponding excess demands $g(p)$ and $g(q)$, also assumed to be distinct. If $g(p)$ is available at the price system q, that is if $q \cdot g(p) \leq 0$, then $g(q)$ must not be available at the price system p, that is $p \cdot g(q) < 0$. We thus have $q \cdot g \cdot (p) \leq 0$ implies $p \cdot g(q) > 0$.

Excess demands of individual agents are mainly of interest in that they determine aggregate excess demands. The aggregate excess demand function simply by defining $f(p)$ as the sum of each agent's $g(p)$. $f : S \to R^n_+$ is obtained simply by defining $f(p)$ as the sum of each agent's $g(p)$.

It is immediate that $f(p)$ is homogeneous, that is $f(tp) = f(p)$ for all positive t, and also that $p.f(p) = 0$ for all p, a property known as Walras's Law. It is also clear that f is continuous at any strictly positive p, for then the value of each agent's endowment must be positive, so that each g is continuous. In fact f is continuous everywhere on S, essentially because at any p there must be some agent whose endowment has positive value.

The revealed preference property, which applies to individual agents' excess demands, does not, however, carry over to aggregate excess demands. If all agents are identical, that is, have identical preferences and endowments, then the property carries over, but it may be that individual agents' excess demands are related in a perverse way, which means that the property is lost in aggregation. If the property does apply in aggregate then we may consider the economy to behave as an individual agent.

The excess demand function f provides us with a simple way to define equilibrium, and various of its properties. A price system p in S is an equilibrium price if $f(p) = 0$. Further, for example, an equilibrium price p is unique if, for any q in S, $f(q) = 0$ implies that $q = tp$ for some positive t. Equilibrium and its properties may be defined without the concept of excess demand, but the concept provides a useful framework for their investigation.

For example, with two commodities (and thus effectively one price and one excess demand function) it is intuitively clear that equilibrium will be unique, and indeed stable, if excess demand is everywhere downward-sloping. More generally, uniqueness and stability are typically investigated using generalizations of this idea of downward-sloping excess demand.

We have noted that aggregate excess demands have the properties of homogeneity, Walras's Law and continuity. It is also important to note what is in effect the converse of this: that any function with these properties may be an excess demand function. This is to say that economic theory places no restrictions on excess demand functions other than these three properties.

This property was first investigated by Sonnenschein (1973), but the most powerful result is due to Debreu (1974), who showed that

there is an economy with precisely n agents which generates any continuous homogeneous excess demand function obeying Walras's Law, at least on the set of strictly positive prices. This is proved by first decomposing f into n individual excess demand functions, each having the revealed preference property, and then showing that these individual functions are the result of preference maximization subject to a wealth constraint.

Indeed, if f is restricted to being twice differentiable then there is an economy with n agents, each with homothetic preferences, which generates f as its excess demand function, again on the set of positive prices (Mantel 1976). This is shown by creating indirect utility functions for each agent and then using Roy's Identity to obtain the excess demand functions. However, the extension of this result, and also Debreu's result, to the whole of S, rather than only the set of strictly positive prices, remains open.

Not surprisingly, these results imply that the set of equilibrium prices has little structure. Indeed, it may be any non-empty compact subset of S (Mas-Colell 1977). Thus without further restrictions we can say very little about the set of equilibria.

Bibliography

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1(1): 15–21.

Hicks, J.R. 1946. *Value and capital*, 2nd ed. Oxford: Clarendon Press.

Mantel, R. 1976. Homothetic preferences and community excess demand functions. *Journal of Economic Theory* 12(2): 197–201.

Mas-Colell, A. 1977. On the equilibrium price set of an exchange economy. *Journal of Mathematical Economics* 4(2): 117–126.

Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6(4): 345–354.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Definitive edn, Lausanne: Corbaz, 1926. Trans. W.Jaffé as *Elements of Pure Economics*. London: George Allen & Unwin, 1954.

Excess Volatility Tests

Stephen F. LeRoy

Abstract

Stock prices frequently undergo big changes that do not coincide with commensurate changes in fundamentals (earnings, dividends, interest rates). Shiller and LeRoy and Porter formalized the idea that price volatility is excessive relative to fundamentals by deriving the implications for price volatility of the hypothesis that stock prices equal the present value of discounted dividends. Subsequent discussion focused on the extent to which their results were subject to econometric problems. Also, analysts observed that the adopted version of the present-value model presumed constant discount rates, as would be the case under risk neutrality. This possibly biases the results.

Keywords

Asset pricing; Equity premium puzzle; Excess volatility tests; Payoff volatility; Present value; Price volatility; Price–consumption ratio; Risk aversion; Risk neutrality; Samuelson, P.; Vector autoregressions

JEL Classifications

C5

It has been known for many years that stock prices frequently undergo big changes that do not coincide with commensurate changes in prospective corporate earnings or dividends or in variables that can readily be connected to discount factors, such as interest rates (see for example Cutler et al. 1989). The best-known episode occurred on 19 October 1987, when stock prices dropped around the world – by 22 per cent in the United States – in the complete absence of news about fundamentals. Such events appear to conflict with

finance theory: if stock prices equal the present value of future expected dividend streams, changes in prices should be attributable to news about dividends or discount factors. However, it is difficult to draw reliable conclusions about price volatility from individual episodes, if only because there is no obvious way to evaluate statistical significance.

Is Price Volatility Systematically Excessive?

The question is not whether stock price volatility appears to exceed that justified by fundamentals in individual episodes, but whether it does so *systematically*. The latter question was first addressed by Shiller (1981) and LeRoy and Porter (1981). These papers, written independently and approximately contemporaneously, derived the existence of bounds on the volatility of prices and returns that are implied by the present-value relation. They found excess volatility. However, the papers made this point using different analytical methods. Shiller observed that, if stock prices equal the expectation of summed discounted dividends, then stock price volatility should be bounded above by the volatility of what he called ex-post rational stock prices, defined as actual summed discounted dividends. Ex-post rational prices p_t^*, he pointed out, obey the relation

$$p_t^* = \beta\left(p_{t+1}^* + d_{t+1}\right), \tag{1}$$

where β is a discount factor (assumed constant). From (1), a time series for p_t^* can be constructed by a backward recursion, at least given an initial condition. Shiller constructed graphs of p_t^* and p_t and argued from visual inspection of these graphs that the former was much smoother than the latter, proving that volatility is excessive. Since Shiller did not specify the model assumed to generate the data, he had no way to evaluate statistical significance.

LeRoy and Porter, in contrast, adopted a model-based analytical procedure. They specified that dividends, stock prices and any auxiliary variables that serve as predictors for future dividends are generated by a linear vector autoregression (to use a term that was not yet in vogue). They proved that a certain function of the variance of stock price and the variance of stock payoffs can be derived from the parameters describing the bivariate autoregression for dividends and prices. The reason both price volatility and dividend volatility enter the expression is that, if the auxiliary variables are accurate predictors of future dividend innovations, then price volatility will be high but payoff volatility will be low. The opposite will be the case if the auxiliary variables do not give accurate predictions of future dividends.

Using this result, LeRoy and Porter constructed a joint test of price volatility and payoff volatility from a bivariate model for dividends and prices. It was unnecessary to estimate the forecasting power of the auxiliary variables, or even to specify them. LeRoy and Porter conducted this test and reported a confidence interval based on the asymptotic distribution of the coefficients of the bivariate process for dividends and prices. They found excess volatility, but it appeared to be of borderline significance. See LeRoy (1989) for a fuller, but still brief, summary of the variance-bounds tests in the context of the efficient capital markets literature.

Econometric Problems

Both Shiller's and LeRoy and Porter's procedures had econometric problems (for a survey of these problems, see Gilles and LeRoy 1991). These problems were serious enough to invalidate the results, in the opinion of some. Discussion focused on Shiller's paper. Kleidon (1986a, b) and Flavin (1983) pointed out that, while the present-value model implied that the unconditional variance of p_t^* exceeded that of p_t, one would expect the variance of p_t^* conditional on its neighbours p_{t+j}^* to be lower than the corresponding conditional variance of p_t. This is so because p_t^* is much more highly autocorrelated than p_t, as is evident from the absence of an error term in (1). In visually evaluating the volatility of p_t^* and p_t it is not easy to distinguish unconditional

from conditional volatility. This is a major drawback of Shiller's procedure. Kleidon computed simulations of p_t^* and p_t in which the present-value model was true by construction, and argued that they looked much the same as the actual data.

LeRoy and Porter's procedure had the drawback that the assumed linear process for dividends and prices implies that these are stationary in level. That being so, some sort of trend correction to remove the upward trend in both variables must be imposed, and LeRoy and Porter did so. Trend-correction algorithms can easily distort the time-series properties of variables, and that may have happened in this case. It is possible that LeRoy and Porter's finding that excess price and return volatility is only marginally significant statistically reflects difficulties with the trend correction.

Interpreting Excess Volatility

Analysts questioned the interpretation of the variance-bounds tests along other lines. They pointed out that the volatility implications of the present-value relation were just repackaged versions of the fundamental implication of the present-value relation (plus the assumption of rational expectations) that stock returns should be serially uncorrelated. If so, why do direct tests of the return orthogonality implication of the present-value relation tend to accept the null, whereas the variance bounds tests appear to reject it? Part of the resolution of this apparent contradiction is that the evidence on return uncorrelatedness began to look less favourable to the null hypothesis in the late 1980s (see for example Campbell and Shiller 1988). Another possibility, discussed by LeRoy and Steigerwald (1995), is that the volatility tests are more powerful than the return non-autocorrelatedness tests under whatever alternative hypothesis generated the data.

Risk Aversion

Discussion of these issues ended fairly suddenly in the mid-1990s. This happened mostly because of a growing realization that the hypothesis being tested required an assumption of risk neutrality: in general, stock prices equal the discounted value of expected dividends (when the expectation is taken under the natural probabilities) with a non-stochastic discount factor only if agents are risk neutral). If agents are risk averse there is no reason to presume that either the orthogonality implications or the volatility implications of the present-value relation will be satisfied. This dependence on risk neutrality had not been brought out clearly in the major papers developing the orthogonality implications of market efficiency. For example, Samuelson's otherwise superb paper (1965) developing the connection between martingale models and the present-value relation glossed over this point. Similarly, Fama's classic 1970 survey of the efficient capital markets literature observed that 'market efficiency' (meaning, presumably, rational expectations) could be tested only if the analyst committed himself to a particular model specifying how returns are generated, so that the joint hypothesis of efficiency and the assumed model is tested. Despite this, Fama did not go on to observe that the returns model that underlies conventional market efficiency tests took no account of risk aversion. Shortly it became clear that the non-autocorrelatedness of returns would not occur in general if agents are risk averse (LeRoy 1973, 1976; Lucas, 1978).

Initially the dependence of the variance-bounds tests on risk neutrality appeared to be a somewhat abstract point. However, arguments were shortly presented that this might not be so. LaCivita and LeRoy (1981), using a two-state version of Lucas's (1978) tree model, showed that allowing for risk aversion could be expected to increase the predicted volatility of stock prices. Risk-averse agents will try to smooth consumption across time by transferring consumption from low-marginal-utility states to high-marginal-utility states. However, in an exchange economy they cannot do so in the aggregate. The representative agent must consume the aggregate endowment in equilibrium, so prices must counteract preferences. If stock prices are very high when the marginal utility of consumption is low, and vice versa, agents must buy

financial assets when they are expensive and sell them when they are cheap if they are to transfer claims on consumption as desired. This price pattern decreases their desire to do so. If price volatility is high enough, this effect will induce agents to consume the endowment. A related argument was presented by Grossman and Shiller (1981).

Relation to the Equity Premium Puzzle

The excess volatility debate shifted with the arrival of Mehra and Prescott's wellknown paper (1985) on the equity premium puzzle. This paper was relevant to the excess volatility question for several reasons. The Mehra and Prescott paper followed LaCivita and LeRoy in specializing Lucas's tree model to two states, but modified the states so that they described the growth rate of the aggregate endowment rather than its level, as in Lucas and LaCivita and LeRoy. This leads to a tractable model when the representative agent has homothetic utility, as with power utility. A major advantage of Mehra and Prescott's specification is that when consumption growth rates rather than levels are stationary there is no need for trend correction.

Mehra and Prescott imposed drastic simplifications so as to obtain a tractable model. For example, their model did not distinguish among corporate earnings, dividends and aggregate consumption, despite the fact that these variables behave differently. Some analysts expected that Mehra and Prescott's finding that the equity premium is excessive would be reversed when these simplifications were reversed, but that has turned out not to be the case (see, for example, Kocherlakota 1996).

LeRoy and Parke (1992) observed that Mehra and Prescott's framework can be adapted to the investigation of volatility by imposing the assumption that consumption growth rates are independently distributed. This is a special case of the Markov distribution that Mehra and Prescott specified. In that case the ratio of equity value to consumption follows a stationary process. The volatility of that variable depends on how much information agents have about future consumption beyond that contained in present consumption. In the simplest case, when agents have no such information, price is a constant markup of consumption, implying that stock returns have the same volatility as the consumption growth rate. This is true regardless of the degree of risk aversion, implying that the implications of risk aversion for volatility are very different in stationary-growth-rate models than in the stationary-levels models discussed above. This prediction is rejected by the US data: the standard deviation of annual consumption growth is about two per cent, whereas that of annual stock returns is on the order of 20 per cent. In contrast to the equity premium puzzle, which can in principle be resolved with sufficiently high risk aversion, the prediction that equity returns should have standard deviations around two per cent holds for any level of risk aversion.

Even if one accepts that consumption is a geometric random walk, assuming that agents have no information variables for future consumption other than current consumption is unacceptable. If agents do have such information, equity prices will not be a constant markup of consumption. However, LeRoy and Parke showed that in that case the variances of the price–consumption ratio and the return on stock obey a relation similar to that obtained by LeRoy and Porter. They found that the resulting joint test on the volatility of the price–consumption ratio and the volatility of stock returns results in excess volatility for either variable or both.

Most analysts believe that no single convincing explanation has been provided for the volatility of equity prices. The conclusion that appears to follow from the equity premium and price volatility puzzles is that, for whatever reason, prices of financial assets do not behave as the theory of consumption-based asset pricing predicts.

See Also

► Risk Aversion

Bibliography

Campbell, J., and R. Shiller. 1988. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1: 195–228.

Cutler, D., J. Poterba, and L. Summers. 1989. What moves stock prices? *Journal of Portfolio Management* 15: 4–12.

Fama, E. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 283–417.

Flavin, M. 1983. Excess volatility in the financial markets: A reassessment of the empirical evidence. *Journal of Political Economy* 91: 929–956.

Gilles, C., and S. LeRoy. 1991. Econometric aspects of the variance bounds tests: A survey. *Review of Financial Studies* 4: 753–791.

Grossman, S., and R. Shiller. 1981. The determinants of the variability of stock prices. *American Economic Review Papers and Proceedings* 71: 222–227.

Kleidon, A. 1986a. Bias in small-sample tests of stock price rationality. *Journal of Business* 59: 237–261.

Kleidon, A. 1986b. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94: 953–1001.

Kocherlakota, N. 1996. The equity premium: It's still a puzzle. *Journal of Economic Literature* 34: 42–71.

LaCivita, C., and S. LeRoy. 1981. Risk aversion and the dispersion of asset prices. *Journal of Business* 54: 535–547.

LeRoy, S. 1973. Risk aversion and the martingale model of stock prices. *International Economic Review* 14: 436–446.

LeRoy, S. 1976. Efficient capital markets: Comment. *Journal of Finance* 31: 139–141.

LeRoy, S. 1989. Efficient capital markets and martingales. *Journal of Economic Literature* 27: 1583–1621.

LeRoy, S., and W. Parke. 1992. Stock price volatility: tests based on the geometric random walk. *American Economic Review* 82: 981–992.

LeRoy, S., and R. Porter. 1981. The present value relation: Tests based on implied variance bounds. *Econometrica* 49: 555–574.

LeRoy, S., and D. Steigerwald. 1995. Volatility. In *Handbook of finance*, ed. R. Jarrow, V. Maksimovic, and W. Ziemba. Amsterdam: North-Holland.

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

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

Samuelson, P. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.

Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.

Exchange

Robert B. Wilson

Abstract

Economic studies of exchange examine processes in which agents obtain gains from trade, e.g. bilateral bargaining and contracting, auctions, and multilateral markets. Prominent descriptive theories include idealized versions that assume agents simply respond to prices that clear markets. Realistic versions recognize effects of procedural rules and strategic behavior, and various impediments such as incomplete or unenforceable contracts, insufficient markets, imperfect information, and incomplete observability. Normative versions design market procedures, contract forms, and settlement rules that strengthen incentives and promote efficient outcomes.

Keywords

Adverse selection; Arbitrage; Asymmetric information; Auctions; Bargaining; Barter; Bundling; Coalitions; Collusion; Commitment; Competition; Complementary goods; Complete information; Cores; Decentralization; Double auctions; Edgeworth, F. Y.; Efficiency prices; Efficient allocation; Entry; Excess demand and supply; Exchange; Existence of equilibria; Expected utility; Fixed-point problems; Game theory; Impatience; Implicit contracts; Incentive compatibility; Incomplete contracts; Incomplete information; Incomplete observability; Incomplete markets; Information aggregation and prices; Inside information; Labour market contracts; Marginal rates of substitution; Market design; Mechanism design; Monopoly rents; Moral hazard; Multiple equilibria; Nonlinear pricing; Optimal contracts; Preference falsification; Price discrimination; Principal and agent; Prisoner's dilemma; Private information; Product

differentiation; Quality differentiation; Rational behaviour; Rational expectations; Reputation; Resale markets; Reservation prices; Revelation principle; Risk sharing; Sealed-bid auctions; Sequential rationality; Shapley value; Signalling; Specialization; Strategic behaviour; Temporary equilibrium; Terms of trade; Uniqueness of equilibrium; Vickrey, W. S.; Walras, L.; Walras's Law; Walrasian exchange model

JEL Classifications
D5

The scope of economics includes allocation of scarce resources. Allocation comprises production and exchange, according to a division between processes that transform commodities and those that transfer control. For production or consumption, exchange is essential to efficient use of resources. In production it allows decentralization and specialization; and, for consumption, agents with diverse endowments or preferences require exchange to obtain maximal benefits. Voluntary exchange involves trading bundles of commodities or obligations to the mutual advantage of all parties to the transaction. If two agents have differing marginal rates of substitution, then there exists a trade benefiting both. The advantages of barter extend widely, for example, to trade among nations and among legislators ('vote trading'), but here it suffices to emphasize markets with enforceable contracts for trading private property unaffected by externalities, and with money as a medium of exchange.

In a market economy using money or credit, terms of trade are usually specified by prices denominated in money. Besides purchases at prices posted by producers and distributors, exchange occurs in bargaining, auctions, and other contexts with repeated or competitive offers. In institutionalized 'exchanges' for trading commodities, brokers offer bid and ask prices; and, for trading financial instruments, specialists cross buy and sell orders and maintain markets continually by quoting bid and ask prices and trading for their own accounts.

Theories of Exchange

Records of transaction prices and quantities are the raw data of many empirical studies of economic activity, and explanation of these data is a main purpose of economic theory. Theories of exchange attempt to predict the terms of trade and the resulting transactions from the market structure and the agents' attributes, such as endowments, productive opportunities, preferences, and information. Also relevant are the markets accessible, the trading rules used, and the contracts available. These may depend on property rights, search or transaction costs, and on events observable or verifiable to enforce contracts. If a particular trading rule is used, such as an auction, then it specifies the actions allowed each agent in each contingency, and the trades resulting from each combination of the agents' actions. These features are the ingredients of experimental designs to test theories, and they motivate models used for empirical estimates of market behaviour. Normative considerations are also relevant, so welfare analyses study the distributional consequences of alternative trading procedures and contracts.

Most theories hypothesize that each agent acts purposefully to maximize its (expected utility of) gain from trade. Some behaviour may be erratic, customary, or reflect dependency on a status quo, but experimental and empirical evidence substantially affirms the hypothesis of 'rational' behaviour, at least in the aggregate. Although more general theories are available, the main features are explained by preferences that are quite regular, as assumed here: monotone, convex, and possibly allowing risk aversion and impatience.

Typically there are many efficient allocations of a fixed endowment, since any allocation that equates agents' marginal rates of substitution is efficient. In the case of risk sharing, for example, an allocation is efficient if all agents achieve the same marginal rates of substitution between incomes in every two states. The distribution of endowments among agents evidently matters, however, and a major accomplishment has been the identification of a small set of salient efficient allocations. Named for Léon Walras, this set is a

focus of nearly all theories, in the sense that other allocations are explained by departures from the Walrasian model.

An allocation is Walrasian if it is obtainable by trading at prices such that it would cost each agent more to obtain a preferable allocation. That is, items are bought at uniform prices available to all, and each agent chooses a preferred trade within a budget constraint imposed by the values of goods bought and sold. If markets are complete, then a Walrasian allocation is necessarily efficient, because another allocation preferred by every agent would cost each agent more at the current prices and therefore more in total, which cannot be true if the preferred allocation is a redistribution of the present one. Conversely, each efficient allocation is Walrasian without further trade, since the agents' common marginal rates of substitution serve as the price ratios. The basic formulation considers trade for delivery in all future contingencies, but refined formulations elaborate the realistic case that markets reopen continually and trade is confined to a limited variety of contracts for immediate and future delivery, possibly contingent on events.

Sufficient conditions for Walrasian allocations to exist have been established. Mainly these require that agents' preferences are convex and insatiable, and that each agent has an endowment sufficient to obtain a positive income. For 'most' economies the number of Walrasian allocations is finite, but uniqueness requires strong assumptions on substitution and income effects. Walrasian allocations and prices for a specific model can be computed by solving a fixed-point problem, for which general methods have been devised. The task is complex (for example, linear models with integer data can yield irrational prices) but an important simplifying feature is that Walrasian prices depend only on the distribution of agents' attributes, and in particular only on the aggregate excess demand function. Essentially, any continuous function satisfying Walras's law (at any prices the value of excess demand is zero) and homogeneity in prices is the excess demand for some economy.

The key requirement for a Walrasian allocation is that each agent's benefit is maximized within its budget imposed by the assigned prices, and that markets clear at those prices. However, complete exploitation of all gains from trade may be precluded by incomplete markets, pecuniary externalities (such as absence of necessary complementary goods), insufficient contracts, or strategic behaviour. If producers with monopoly power restrain output to elevate prices, or practise any of the myriad forms of price discrimination, then the resulting allocation need not be Walrasian. Much discrimination segments markets via quality differentiation or bundling, but equally common is discriminatory pricing of the various conditions of delivery (for example, spatial, temporal, service priority) or if purchases can be monitored and resale markets are absent, by nonlinear pricing of quantities (for example, quantity discounts, loyalty programmes, two-part and block-declining tariffs).

The Walrasian model of exchange is substantially defined by the absence of such practices affecting prices. It also relies on a fixed specification of markets, agents, products, and contracts. The theory of economies with large firms having power to influence prices and to choose product designs is significantly incomplete. The deficiencies derive partly from inadequate formulations, and partly from technical considerations – characterizations and even the existence of equilibria depend on special structural features. For example, the simplest models positing simultaneous choices of qualities and prices by several firms lack equilibria; models with sequential choices encounter similar obstacles but to lesser degrees. In addition, if lump-sum assessments imposed on customers are precluded, then recovering firms' large fixed costs may require nonlinear pricing or price discrimination.

Market clearing is also essential to the Walrasian model since prices are determined entirely by the required equality of demand and supply, including inventories in dynamic contexts. In contrast, successive markets with overlapping generations of traders need not clear 'at infinity'. Such markets can exhibit complicated dynamics even if the underlying data of the economy are stationary. Similarly, continually repeated markets where buyers and sellers arrive

at the same rate that others depart after completing transactions (for example, real estate) admit non-Walrasian prices or may have persistent excess on one side of the market if search or dispersed bargaining prevents immediate clearing.

Factors Affecting Exchange

When it is that among the feasible allocations the best prediction might be one of the Walrasian allocations, has been answered in several ways.

Competition is the first answer. On the supply side, for instance, with many sellers each one's incentive to defect from collusive pricing arrangements is increased. Absent collusion, if prices reflect total supplies offered on the market and each seller chooses its optimal supply in response to anticipations of other sellers' supplies, then each seller's optimal percentage profit margin declines inversely with the number of sellers offering substitutes. Price discrimination, such as nonlinear pricing, is inhibited if there are many sellers, resale markets are available, or customers' purchases are difficult to monitor. Absent capacity limitations, direct price competition among close or perfect substitutes erodes profits since undercutting is attractive. Although these conclusions are weakened to the extent that buyers incur search or switching costs, easy entry incurring low sunk costs remains important to ensure that markets are contestable and monopoly rents are eliminated. Monopoly rents are often substantially dissipated in entry deterrence, price wars, and other competitive battles to retain or capture monopoly or oligopoly positions. This is true both when entrants bring perfect substitutes and also more generally, since entrants tend to fill in the spectra of quality attributes and conditions of delivery.

Arbitrage is important in markets for commodities with standardized qualities, especially financial assets and derivatives such as options. To the extent that the contingent returns from one asset replicate those from a bundle of other assets, or from some trading strategy, its price is linked to the latter. Importantly, repeated opportunities to trade contingent on events enable a few securities to substitute for a much wider variety of contingent contracts.

One form of the competitive hypothesis emphasizes that each subset of the traders can redistribute their endowments among themselves. For example, a seller and those buyers who purchase from him are a coalition redistributing their resources among themselves. A core allocation is such that no coalition can redistribute its endowments to advantage every member. The core allocations include the Walrasian allocations. A basic result first explored by F.Y. Edgeworth establishes that as the economy is enlarged by adding replicates of the original traders, the set of core allocations shrinks to the Walrasian allocations. Deeper analyses of core allocations take account of agents' private information, but in this context the relation to Walrasian allocations is tenuous.

Another view emphasizes that in an economy so large that each agent's behaviour has an insignificant effect on the terms of trade, every trader's best option is to maximize its gain from trade at the prevailing prices. For example, any one trader's potential gain from behaviour that influences prices becomes insignificant as the set of traders expands, provided the limit distribution is 'atomless'. Similar results obtain for various models of markets with explicit price formation via auctions or bilateral bargaining. Generally, an efficient allocation is necessarily Walrasian if each agent is unnecessary to attainment of others' gains from trade. An idealized formulation considers an atomless measure space of agents in which only measurable sets of agents matter and thus the behaviour of each single agent is inconsequential. In this case the Walrasian allocations are the only core allocations. Similarly, a Walrasian allocation results from the Shapley value in which each agent shares in proportion to his average marginal contributions to randomly formed coalitions.

Structural features of trading processes suggest alternative hypotheses. Matching problems (for example, workers seeking jobs) admit procedural rules that with optimal play yield core allocations, and for a general exchange economy an appropriately designed auction yields a core allocation. Other games have been devised for which optimal strategies of the agents result

in a Walrasian allocation. Continual bilateral bargaining among dispersed agents with diverse preferences, in which agents are repeatedly matched randomly and one designated to offer some trade to the other, also results in a Walrasian allocation from optimal strategies. In a related vein, several methods of selecting allocations create incentives for agents to falsify reports of their preferences, but if they do this optimally then a Walrasian allocation results. Quite generally, any process that is fair in the sense that all agents enjoy the same opportunities for net trades yields a Walrasian allocation. In one axiomization, some signal is announced publicly and then based on his preferences each agent responds with a message that affects the resulting trades: if a core allocation is required, and each signal could be the right signal for some larger economy, then the signal must be essentially equivalent to announcement of a Walrasian price to which each agent responds with his preferred trade within his budget specified by the price.

Traders' impatience can also affect the terms of trade. In the simplest form of impatience, agents discount delayed gains from trade. Dynamic play is assumed to be sequentially rational in the sense that a strategy must specify an optimal continuation from each contingency – this strong requirement severely restricts the admissible equilibria. For example, if a seller and a buyer alternate proposing prices for trading an item, then in the unique equilibrium trade occurs immediately at a price dependent on their discount rates. As the interval between offers shrinks, the seller's share of the gains from trade becomes proportional to the relative magnitude of the buyer's discount rate; for example, equal rates yield equal division. Extensions to multilateral contexts produce analogous results. A monopolist with an unlimited supply selling to a continuum of buyers might plausibly extract favourable terms, but actually, in any equilibrium in which the buyers' strategies are stationary, as the interval between offers shrinks the seller's profit disappears and all trade occurs quickly at a Walrasian price. Similarly, a durable-good manufacturer lacking control of resale or rental markets has an incentive to increase the output rate as the production period shrinks or to pre-commit to limited capacity. This emphasizes that monopoly power depends substantially on powers of commitment stemming from increasing marginal costs, capacity limitations, or other sources. However, impatience and sequential rationality can produce inefficiencies in product design, since then a manufacturer may prefer inferior durability, or in market structure, since a seller may prefer to rent rather than sell durable goods.

Complete information is a major factor justifying predictions of Walrasian prices. Many theories predict Walrasian outcomes when there is complete information and agents have symmetric trading opportunities, but incomplete information often produces departures from the Walrasian norm.

Although information may be productively useful, in an exchange economy the arrival of information may be disadvantageous to the extent that risk-averse agents forgo insurance against its consequences. A basic result considers an exchange economy that has reached an efficient allocation before some agents receive further private information, and this fact is common knowledge: the predicted response is no further trade, though prices may change.

Each efficient allocation has 'efficiency prices' that reflect the marginal rates of substitution prevailing – in the Walrasian case all trades are made at these prices. They summarize a wealth of information about technology, endowments and preferences. Prices and other endogenous observables are therefore not only sufficient instruments for decentralization but also carriers of information. If information is dispersed among agents then Walrasian prices are signals, possibly noisy, that can inform agents' trading. Models of 'temporary equilibrium' envision a succession of markets, in each of which prices convey information about future trading opportunities. 'Rational expectations' models assume that each agent maximizes an expected utility conditioned on both his private information and the informational content of prices. In simple cases prices are sufficient statistics that swamp an agent's private information. In complex real economies the informational content of prices may be elusive; nevertheless, markets are affected by inferences from prices (e.g. indices of stock and wholesale prices)

and various models attempt to include these features realistically. Conversely, responses of prices to events and disclosures by firms are studied empirically.

The privacy of each agent's information about his preferences and endowment affects the realized gains and terms of trade. In some cases the relative prices of 'qualities' provide incentives for self-selection. An example is a product line comprised of imperfect substitutes, in which price increments for successive quality increments induce customers to select according to their preferences. Several forms of discrimination in which prices depend on the quality (for example, the time, location, priority or other circumstances of delivery) or, if resale is prevented, the quantity purchased, operate similarly.

Absence of the relevant contingent contracts is implicitly a prime source of inefficiencies and distributional effects. Trading may fail if adverse selection precludes effective signalling about product quality: without quality assurances or warranties, each price at which some quality can be supplied attracts sellers offering lesser qualities. Investments in signals, possibly unproductive ones, that are more costly for sellers supplying inferior qualities induce signal-dependent schedules in which the price paid depends on the signal offered. For example, to signal his ability a worker may over-invest in education or work in a job for which he would be underqualified on efficiency grounds. If buyers make repeat purchases based on the quality experienced from trying a product then the initial price itself, or even dissipative expenditures such as uninformative advertising, can be signals used by the seller to induce initial purchases.

Principal–agent relationships in which a risk-averse agent has superior information or his actions cannot be monitored completely by the principal require complex contracts. For example, in a repeated context with perfect capital markets and imperfect insurance, the optimal contract provides the agent with a different reward for each measurable output, and the total remuneration is the accumulated sum of these rewards. Contracting is generally affected severely by limited observation of contingencies (either events or actions relevant to incentives) and in asymmetric relationships nonlinear pricing is often optimal. For example, insurance premia may vary with coverage to counter the effects of adverse selection or moral hazard.

Labour markets are replete with complex incentives and forms of contracting, partly because workers cannot contract to sell labour forward and partly because labour contracts substitute for imperfect loan markets and missing insurance markets (for example, against the risk of declining productivity). Workers may have superior information about their abilities, technical data, or effort and actions taken; and firms may have superior information about conditions affecting the marginal product of labour. Incentives for immediate productivity may be affected by conditioning estimates of ability on current output, or by procedures selecting workers for promotion to jobs where the impact of ability is multiplied by greater responsibilities. The complexity of the resulting incentives and contracts reflects the multiple effects of incomplete markets and imperfect monitoring.

In the context of trading rules that specify price determination explicitly, analyses of agents' strategic behaviour emphasize the role of private information. The trading rule and typically the probability distribution of agents' privately known attributes are assumed to be common knowledge; consequently, formulations pose games of incomplete information. An example is a sealed-bid auction in which the seller awards an item to the bidder submitting the highest price. Suppose each bidder observes an imperfect estimate that is independently and identically distributed (i.i. d.) conditional on the unknown value of the item. With equilibrium bidding strategies, as the number of bidders increases the maximal bid converges in probability to the expectation of the value conditional on knowing the maximum of all the samples; for the common distributions this implies convergence to the underlying value. Alternative auction rules are preferred by the seller according to the extent that the procedures dilute the informational advantages of bidders (for example, progressive oral bidding has this effect) and exploit impatience and risk aversion. Rules can be constructed that maximize the seller's expected revenue: if bidders' valuations are i.i.d. then for the common distributions awarding

the item to the highest bidder at the first or second-highest bid is optimal, subject to an optimal reservation price set by the seller. In such a second-price or oral progressive auction with no reservation price, bidders offer their valuations, so the price is Walrasian.

Another example is a double auction, used in various commodity and financial markets, in which multiple buyers and sellers submit bid and ask prices and then a clearing price is selected from the interval obtained by intersecting the resulting demand and supply schedules. For a restricted class of models, requiring sufficiently many buyers and sellers with i.i.d. valuations, a double auction is incentive efficient, in the sense that there is no other trading rule that is sure to be preferred by every agent; also, as the numbers increase the clearing price converges to a Walrasian price.

The effects of privileged 'inside' information held by some traders have been studied in the context of markets mediated by brokers and specialists, as in most stock and option markets. The results show that specialists' strategies impose all expected losses from adverse selection on uninformed traders. Specialists may further profit from knowledge of the order book and immediate access to trading opportunities.

Private information severely affects bargaining. With alternating offers even the simplest examples have many equilibria, plausible criteria can select different equilibria, and a variety of allocations are possible. In most equilibria, delay in making a serious offer is a signal that a seller's valuation is not low or a buyer's is not high; or the offers made limit the inferences the other party can make about one's valuation. When both valuations are privately known, signalling must occur in some form to establish that gains from trade exist. Typically all gains from trade are realized eventually, but with significant costs of delay. Applications extend beyond purely economic contexts, such as to negotiations to settle a law suit.

In a special case, a seller with a commonly known valuation repeatedly offers prices to a buyer with a privately known valuation: assume that the buyer's strategy is a stationary one that accepts the first offer less than a reservation price depending on his valuation. As mentioned previously for the monopoly context, as the period between offers shrinks the seller's offers decline to a price no more than the least possible buyer's valuation and trade occurs quickly – thus the buyer captures most of the gains. Even with alternating offers, the buyer avoids serious offers if his valuation is high and the periods are short. Thus, impatience, frequent offers, and asymmetric information combine to skew the terms of trade in favour of the informed party.

The premier instance of exchange is the commodity trading pit in which traders around a ring call out bid and ask prices or accept others' offers. These markets operate essentially as multilateral versions of bargaining but with endogenous matching of buyers and sellers: delay in making or accepting a serious offer can again be a signal about a trader's valuation, but with the added feature that 'competitive pressure' is a source of impatience. That is, a trader who delays incurs a risk that a favourable opportunity is usurped by a competing trader. These markets have been studied experimentally with striking results: typically most gains from trade are realized, at prices eventually approximating a Walrasian clearing price, especially if the subjects bring experience from prior replications. However, if complicated 'rational expectations' features are added, then subjects may fail to infer all the information revealed by offers and transactions.

Trading rules can be designed to maximize the expected realized gains from trade, using the 'revelation principle'. Each trading rule and associated equilibrium strategies induce a 'direct revelation game' whose trading rule is a composition of the original trading rule and its strategies; thus in equilibrium each agent has an incentive to report accurately his privately known valuation. In the case that a buyer and a seller have valuations drawn independently according to a uniform distribution, the optimal revelation rule is equivalent to a double auction in which trade occurs if the buyer's bid exceeds the seller's offer, and the price used is halfway between these. Basic theorems establish that private information among buyers and sellers precludes realization of all the gains from trade – but if traders are symmetric, as when a trader might buy or sell depending on the price, then sometimes full efficiency can be attained. Generally, with

many buyers and sellers and an optimal trading rule, the expected unrealized gain from trade declines quickly as the numbers of buyers and sellers increase. Such static models depend, however, on the presumption that subsequent trading opportunities are excluded.

Enforceable contracts facilitate exchange, and most theories depend on them, but they are not entirely essential. Important in practice are 'implicit contracts' that are not enforceable except via threats of discontinuing the relationship after the first betrayal. Similarly, in an infinitely repeated situation, if a seller chooses a product's quality (say, high or low) and price before sale, and a buyer observes the quality only after purchasing, then the buyer's strategy of being willing to pay currently only the price associated with the previously supplied quality suffices to induce continual high quality.

Studies of exchange without enforceable contracts focus on the Prisoner's Dilemma game in which both parties can gain from exchange, but each has an incentive to renege on his half of an agreement. In any finite repetition of this game with complete information the equilibrium strategies predict no agreements, since each expects the other to renege. Infinite repetitions can sustain agreements enforced by threats of refusal to cooperate later. With incomplete information, reputational effects can sustain agreements until near the end. For example, if one party thinks the other might surely reciprocate cooperation then he has an incentive to cooperate until first betrayed, and the other has an incentive to reciprocate until defection becomes attractive near the end. Reputations are important also in competitive battles among firms with private cost information: wars of attrition select the efficient survivors.

In sum, the Walrasian model remains a paradigm for efficient exchange under 'perfect' competition in which equality of demand and supply is the primary determinant of the terms of trade. Further analysis of agents' strategic behaviour with private information and market power elaborates the causes of incomplete or imperfectly competitive markets that impede efficiency, and it delineates the fine details of endogenous product differentiation, contracting, and price formation essential to the application of the Walrasian model.

Game-Theoretic Analysis of Exchange

Studies of exchange rely increasingly on game-theoretic methods. These are useful to study strategic behaviour in dynamic contexts; to elaborate the roles of private information, impatience, risk aversion, and other features of agents' preferences and endowments; to describe the consequences of incomplete markets and contracting limited by monitoring and enforcement costs; and to establish the efficiency properties of the common trading rules. They also integrate theories of exchange with theories of oligopolistic collusion, product differentiation, discriminatory pricing, and other strategic behaviour by producers. Technically, the game-theoretic approach enables a transition from theories of a large competitive economy with a specified distribution of agents' attributes, to theories of an economy with few agents, each having private information and acting strategically to exploit opportunities.

Game-theoretic methods are especially useful in market design, that is, devising trading rules and procedures that yield outcomes that are efficient subject to the limitations imposed by participants' private information and strategic behaviour. Innovative designs are used for auctions of government procurements and privatization of assets (for example, spectrum licences, Treasury securities), for wholesale commodity markets such as electricity and gas among many others, and some markets for transport. Some of these are 'smart markets' in that the allocation is derived from an elaborate optimization that takes account of bids and offers for several commodities and various technical constraints – such as transmission limits and reserve requirements in the case of electricity. With the advent of electronic commerce, innovative designs are also used in some retail markets conducted as auctions.

A salient feature of the developing theory of market design is the role of alternative settlement rules. Unlike the Walrasian rule of settling all transactions at the market clearing prices, these settlements provide incentives for accurate reporting of benefits and costs. Most of these rules are derivatives of one proposed by

W. Vickrey (1961) in which each buyer in an auction pays the highest rejected bid; and more generally (even for public goods), for the allocation he receives each trader is charged for the benefit thereby denied to others. Novel settlement rules that promote 'incentive compatibility' and thus discourage strategic behaviours that impair efficiency are hallmarks of the general theory of 'mechanism design'. Essentially, this approach to exchange replaces the idealized descriptive approach in the Walrasian paradigm with a normative 'engineering' construction of optimal trading and settlement rules, that is, a comprehensive contract that governs participation in the market. Participants' perceptions of the effectiveness of this contract can affect competition for trading volume among alternative market venues.

See Also

- ▶ Adverse Selection
- ▶ Agency Problems
- ▶ Auctions (Theory)
- ▶ Bargaining
- ▶ Contract Theory
- ▶ Cores
- ▶ Efficient Allocation
- ▶ Epistemic Game Theory: Incomplete Information
- ▶ Experimental Economics
- ▶ Fair Allocation
- ▶ Game Theory
- ▶ General Equilibrium
- ▶ Incomplete Contracts
- ▶ Incomplete Markets
- ▶ Matching and Market Design
- ▶ Mechanism Design
- ▶ Perfect Competition
- ▶ Price Discrimination (Theory)
- ▶ Principal and Agent (i)
- ▶ Walras's Law

Bibliography

Akerlof, G. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 89: 488–500.

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.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Ausubel, L.M., P.C. Cramton, and R.J. Deneckere. 2002. Bargaining with incomplete information. In *Handbook of game theory*, vol. 3, ed. R.J. Aumann and S. Hart. Amsterdam: North-Holland.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

Fudenberg, D., and J. Tirole. 1993. *Game theory*. Cambridge, MA: MIT Press.

Gül, F., and A. Postlewaite. 1992. Asymptotic efficiency in large exchange economies with asymmetric information. *Econometrica* 60: 1273–1292.

Gül, F., and H. Sonnenschein. 1988. On delay in bargaining with one-sided uncertainty. *Econometrica* 56: 601–611.

Gül, F., H. Sonnenschein, and R.B. Wilson. 1986. Foundations of dynamic monopoly and the coase conjecture. *Journal of Economic Theory* 39: 155–190.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Hölmstrom, B.R., and P.R. Milgrom. 1986. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55: 303–328.

Hölmstrom, B.R., and R.B. Myerson. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51: 1799–1820.

Kreps, D.M., P.R. Milgrom, D.J. Roberts, and R.B. Wilson. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* 27: 245–252.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 54–71.

Milgrom, P.R. 1979. A convergence theorem for competitive bidding with differential information. *Econometrica* 47: 679–688.

Milgrom, P.R. 2004. *Putting auction theory to work*. Cambridge: Cambridge University Press.

Milgrom, P.R., and N. Stokey. 1982. Information, trade, and common knowledge. *Journal of Economic Theory* 26: 17–27.

Myerson, R.B., and M.A. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.

Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40: 289–303.

Roberts, D.J., and A. Postlewaite. 1976. The incentives for price-taking behavior in large exchange economies. *Econometrica* 44: 115–128.

Roberts, D.J., and H. Sonnenschein. 1977. On the foundations of the theory of monopolistic competition. *Econometrica* 45: 101–113.
Roth, A.E. (ed.). 1995. *Game-theoretic models of bargaining*. Cambridge: Cambridge University Press.
Roth, A.E., and M. Sotomayor. 1990. *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge: Cambridge University Press.
Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.
Satterthwaite, M.A., and S.R. Williams. 1989. The rate of convergence to efficiency in the buyers' bid double auction as the market becomes large. *Review of Economic Studies* 56: 477–498.
Scarf, H. (With T. Hansen.) 1973. *The computation of economic equilibria.* New Haven: Yale University Press.
Schmeidler, D. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48: 1585–1593.
Schmeidler, D., and K. Vind. 1972. Fair net trades. *Econometrica* 40: 637–642.
Smith, V. 1982. Microeconomic systems as experimental science. *American Economic Review* 72: 923–955.
Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40: 549–563.
Sonnenschein, H. 1974. An axiomatic characterization of the price mechanism. *Econometrica* 42: 425–434.
Spence, A.M. 1973. *Market signalling: Information transfer in hiring and related process*. Cambridge, MA: Harvard University Press.
Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37.
Wilson, R.B. 1985. Incentive efficiency of double auctions. *Econometrica* 53: 1101–1116.
Wilson, R.B. 1993. Design of efficient trading procedures. In *The double auction market: Institutions, theories, and evidence*, Santa Fe Institute Studies in the Sciences of Complexity, ed. D. Friedman and J. Rust. Reading: Addison-Wesley.

Exchange Control

Pascal Petit

In a rather narrow sense, we refer to exchange control when monetary institutions (governments, central banks or specialized institutions) impose strictly defined limitations on international transactions or on the exchange of national currency into foreign currency. So exchange control occupies the middle ground between unrestricted convertibility into foreign exchange and the total ban on convertibility which is practised in a number of developing countries and in the socialist countries.

In dealing with balance of payments, these restrictions serve different purposes. The most frequent objectives consist in balancing trade (the export and import of goods) or the current account.

In all instances, exchange control measures aim at preserving the autonomy of domestic policies threatened by trade deficits, foreign debts, or a switch in control of the national productive capital. At all times, the primary purpose of exchange controls, as well as that of the non-convertibility of certain currencies, has consisted in preserving the national autonomy of a country from outside interference. The Sparta of Lycurgus, which in turn inspired Plato's project on non-convertible fiduciary currency, provides us with the first famous example (see Einzig 1962). Here the object was to reduce the possibility of corruption by foreign agents. In times of crisis or war, such measures represent a means to control trade or to guarantee the supply of primary strategic materials. But such practices are relatively recent. Einzig (1934) rightly points out that only after 1917 did the World War I belligerents attempt to control exchange. A return to unlimited convertibility occurred only during the second half of the 1920s, after the return to the gold standard in 1925. But the experience gained during World War I facilitated a quick return to control measures during the crises of the 1930s and especially during World War II.

During the decades following the end of World War II, the purpose of controls consisted in limiting any imbalance which in some countries went hand in hand with the development of trade. Basically, controls are introduced when monetary exchange systems fail to fulfil their role as regulators in the international market.

Due to the diversity in exchange control measures, it is difficult to measure their impact. The role of most controls is to keep the current balance of payments in check. In this sense, controls affect the financing of imports (prior demand, necessary deposits, specific rates of exchange), terms of payment (fixed payment delays for export or imports) or limits to travel spending. The exercise of some form of control is the norm rather than the exception.

In the first half of the 1980s – a period characterized by the liberalization of capital movements – only the United States, Switzerland, Britain (since 1979) and the Federal Republic of Germany (since 1984) allowed the free circulation of capital. Japan, which applied vigorous exchange controls until the end of the 1970s, turned to a gradual liberalization spurred by the US–Japanese negotiations in 1984. While improving convertibility, France, Italy and Belgium have kept certain restrictions relating to the circulation of capital.

The rationale underlying such restrictive practices can first of all be found in the limitations imposed by the two traditional regimes of exchange at fixed and flexible rates. We cannot expect such systems to lead to unique equilibrium rates of exchange (a fantasy already denounced by Joan Robinson in 1937). The purpose of such systems is to prevent the development of strong creditor or debtor positions. From time to time these limitations force the countries in question to inhibit the application of free convertibility regimes.

But this is not the sole reason for restricting exchange. Certain control measures aim to shield the development of domestic industries from foreign competition. Here it is useful to distinguish measures bearing on the exchange of commodities and those having effect on the movement of capital, in order to perceive the different stages in a policy geared towards protecting national economies.

In order to understand the advantages or disadvantages of such measures, an attempt will be made to define the limits to exchange regimes and the diversity in control measures.

Exchange Controls as a Reaction to Limitations of the Two Classical Exchange Regimes

For some countries, standard exchange regimes lead to long-lasting imbalances (excessively high debts, inflation triggered by depreciation of the national currency) which can inhibit the autonomy of domestic policies, and so may lead to control measures. The character of these disturbances varies according to the current exchange regime, as demonstrated by the experience of the principal market economies since 1945; at the beginning of the 1970s, these market economies passed from a system of fixed exchange to a system of floating exchange.

In a system of fixed exchange, the defence of parities leads to vast movements of currency reserves by the central banks, inciting speculative movements by private capital, in turn staking on the sudden realignment of parities. The destabilizing effect produced by credit balance or imbalance necessitates parity adjustments or drastic economic measures.

In a system of flexible exchange, where financial markets are largely integrated, domestic economic policies are bounded in terms of prices and rates of interest. This leads to instability in the adjustment of the balance of payments. On the one hand, the increase in the cost of imports in the case of depreciation of the national currency stimulates inflationary pressures and inhibits the re-establishment of foreign debt. On the other hand, the integration of financial markets forces national real interest rates to align with international levels. Speculation then is a measure of the ability of an economic policy to accommodate any pressures on prices and on rates of interest.

The fixed rate regime followed by the flexible rate experience demonstrated the instability generated by free convertibility regimes for medium sized countries, faced with either the risk associated with the issue of a standard currency, i.e. the dollar, or with the erratic speculative movements resulting from its decline.

The gold convertibility of the American currency (at a fixed price of 35 dollars an ounce) constituted one of the pillars of the system of fixed exchange inaugurated at Bretton Woods in 1944. The United States then held eighty per cent of the world gold stock. From the 1960s onwards, military expenses abroad and investment abroad, as well as the balance of trade deficit, led to a sharp decrease in gold reserves and to accumulation by European and Japanese central banks. In early 1971, gold reserves corresponded to only a third of foreign holdings in dollars and the flight of capital increased. In March 1973, an

intensification in speculative movements – encouraged by the announcement by the US Secretary of State of the future total liberalization of capital movements – forced the central banks to endorse the general free floating of currencies under the auspices of the International Monetary Fund.

For current critics of the system of fixed parities (Johnson 1969; Mundell 1968), only a system of flexible exchange could allow for a degree of autonomy and stability in domestic policies. Yet after more than a decade, the system of floating exchange seems conducive to remarkable instability and interdependence between national policies.

With the change in the hegemonic role of the dollar (see Parboni 1981), the monetary uncertainty which characterized the beginning of the 1970s soon led to inflationary pressures. In terms of domestic policy, it was manifest in the large price increase of raw materials together with the explosion of wage conflicts towards the end of the 1960s. In terms of economic policy, inflation represented the main preoccupation of the decade. In general terms, the new symmetry between internationally used currencies made 'financial markets very sensitive to waves of anticipations, polarised alternatively by political and financial events' (Aglietta 1984).

These arbitrary movements reinforced constraints upon national monetary policies. The development of a whole literature using game theory (under the influence of Hamada, 1974) highlights the development of this interdependence.

When confronted with such uncertain situations, a country whose currency has no international role tends to preserve the autonomy of its domestic policy either by participating in the erection of monetary blocs, or by limiting the convertibility of its currency.

The creation in 1977 of the European Monetary System in order to reduce pressures on the monetary policies of the European countries induced by variations in American interest rates, represents a reaction of the first type. In fact, during the 1960s Mundell and McKinnon advocated the creation of a zone of fixed parities in a system of flexible exchange.

The permanency of exchange control measures (when the growth in trade by nature tends towards the dismantling of such measures) together with recommendations by economists such as Tobin (1978) in favour of some control on short-term movements of capital, represents a reaction of the second type.

In exchange regimes which permit the development of important imbalances or which cannot cope with sudden crises of confidence, restriction on free convertibility appears as one of the only means at the disposal of an isolated country which aims at preserving a degree of autonomy in the elaboration of its economic policy. No exchange regime can dispose of the plurality of currencies and the predominance of one of these currencies, such as the dollar. This is a continuous source of conflict (Brunhoff 1986). The attempt by monetarists such as Friedman to present a flexible exchange system as a means to merge the diversity of national currencies into one neutral international currency turned out to be irrelevant.

It is against this background, and in the face of imbalances in terms of employment or external payments, that the relative maintenance of control measures has to be seen, especially since in the postwar years organizations such as the International Monetary Fund or OECD have concurrently been aiming at the liberalization of trade and the convertibility of currencies.

This retention of restrictive practices is all the more remarkable in view of the fact that since the postwar period trade has been more multilateral and diversified than at any other time. In the more segmented world trade system illustrated by the three trade blocs – dollar, pound, gold – in the late 1930s, or in a trade system involving economies at very different stages of development, the limitations to free convertibility appear even more obvious. In fact, there is a strong relationship between circulation of commodities and means of payment. This link plays a decisive role in the evolution in exchange practices.

The Double Nature of Exchange Controls

Commercial and financial aspects of international economic relations are largely dissociated in a tradition which ex-post can be qualified as monetarist-inspired. Bergsten and Williamson (1983) attribute this bias mainly to the implicit (and erroneous) hypothesis of an automatic and regulating adjustment of parities. This dichotomous approach to international economic relations is also apparent in the distribution of roles between international institutions in the immediate postwar period. While the role of GATT consists in reducing barriers to trade, the role of the IMF consists of reducing limitations on the free convertibility of currencies (according to Article 1 of its 1944 Statute).

Yet it is clear that these two aspects – commercial and monetary – of international relations are linked. The orientation and the level of commercial exchange is influenced by financial conditions and conditions of payment. Reciprocally, the development of commercial exchange stimulates the extension of banking networks and financial innovation. In the case of so-called invisible trade, the two areas tend to merge.

This is evident in the case of exchanges of investment revenues realized externally. But it is also applicable to specialized services calling for the establishment of subsidiaries abroad (in the case of financial and insurance activities, consultancies, chambers of commerce etc.). The freedom of capital movements and the right to settle are thus pre-conditions for more advanced specialization by developed countries in the exchange of services. But these specialized services also play a strategic role in commercial exchanges: they are mainly produced by multinational firms and mainly used by other multinational firms which occupy a dominant position in world trade (see Clarimonte and Cavanagh 1984).

The liberalization of invisible trade is therefore not automatic since it implies ample liberalization of capital movements. Since the Tokyo Round the United States has led a campaign within GATT for the liberalization of invisible trade. OECD, which already in the 1950s defined a liberalization code for invisible trade, has noted its ineffectiveness, particularly in the case of the right to do business in a foreign country. Opposition to such policies is manifest in developing countries, where control over the exchange of services is considered strategic (in the early 1980s, liberalization in the area of transport is still a very touchy question).

To this basic complementarity between exchanges of goods and capital movements must be added an interdependence between the different forms of control on monetary exchange and on commercial transactions. Monetary authorities and private agents use the former in order to complement or to thwart the latter. To this end, juggling terms of payment goes hand in hand with fictional or falsified commercial transactions in order to avoid measures that control exchange. Quotas imposed on commercial transactions are sometimes aimed principally at reducing monetary transfers (as illustrated by French experience in the 1930s according to Einzig 1934).

So it is possible like Krueger (1978), Bhagwati (1978) and McKinnon (1979) to consider together all the different forms of controls on commercial and monetary relations between economies in order to define the different forms of foreign trade ‘regimes’. On the basis of national studies of practices which inhibit the freedom of trade (and exchange) in eleven developing countries, Krueger (1978) and Bhagwati (1978) have identified five (not necessarily consecutive) stages. The first stage is characterized by the establishment of generalized and undifferentiated control over imports; this kind of control often follows an unbearable balance of payments deficit. During the second stage there is a large differentiation in the measures of control according to the way in which imports are utilized. A third stage implements a reduction of direct controls (or quotas) together with a sharp devaluation of the national currency. Following this stage there is either a return to the situation described in the second stage or an attempt to liberalize trade by replacing some quotas by tariffs. Finally, the fifth stage is characterized by free convertibility of the national currency for current transactions, which now are only subject to customs duties. So it is only during

this last stage that the commercial and monetary aspects of foreign relations are dissociated.

To this must be added the possibility of controlling capital movements (and therefore of controlling exchange) that influence the trade dynamics mentioned above in respect to developed countries.

Krueger and Bhagwati measure the degree of liberalization of trade according to the disappearance of all types of quotas – despite an increase in customs duties – but nevertheless stress the damaging effects of restrictive practices. Such a position raises two paradoxes.

First, according to a classical theorem in international economics, quotas and customs dues have equivalent effects. Second, why are practices that are hardly 'optimal' so widespread? Krueger and Bhagwati clearly stress that conditions required by the theorem of equivalence so to speak never coincide. The distribution of import licences in the case of quotas, with endogenous effects on supply and demand, modifies resource allocation resulting from custom taxation. The authors also stress the great variety of distributive criteria which lead to distortions. Finally we must consider the question of the rationale for control measures which, according to McKinnon (1979), is hardly considered in the theoretical literature or in case studies. McKinnon brings up the question of political power created by controlling foreign trade in developing countries. Such discretionary power in the distribution of licenses favours 'clientisme', especially if the informal character of domestic activities limits the possibility for internal control. But the importance of such controls is principally derived from the possibility of selecting productive activities. An import licence ensures the viability of an enterprise, in turn giving it access to other limited resources such as capital. The absence or weakness of a financial market, capable of directing sufficient funds toward activities having priority, is according to McKinnon (1979) an essential factor in the origin of trade control practices. The volatility of domestic capital is a sign of this weakness in the financial market.

This explanation for protectionist policies appears to be applicable to the analysis of exchange controls practised in developed countries. The weakness or narrowness of financial markets seems to be one of the major causes for restrictions on the free circulation of currencies. This is indeed suggested by the history of exchange controls.

Past and Present Reasons for Exchange Control Policies

It is seldom asked what imperatives lead to the application of such obviously inconvenient exchange control practices, which simultaneously contain vast opportunities for fraud, potentially ad hoc measures and weak global coherence. Indeed, it is too often a question of faith on the part of opponents of exchange control, who blame a pernicious propensity to bureaucratize to explain the choice of easy direct controls in the place of rigorous but unpopular policies. Their position is not without foundation. But it remains secondary in the face of the risks involved in massive movements of capital. The financial pressures mentioned by McKinnon (1979) do not constitute a marginal phenomenon. No financial market is safe from the flight of capital, feeding itself rapidly to the extent of changing the policies pursued or putting an end to the free circulation of capital. It has been stressed that standard exchange systems do not have the stabilizing effects necessary to prevent speculative moves. On the contrary, there is clear evidence that orthodox exchange systems allow for the maintenance of 'over-evaluation' or 'under-evaluation' of parities. It is within these limits to exchange regimes that major necessary conditions for restrictive practices must be found.

The importance of financial markets and the voluntarist and innovative character of current economic policies constitute sufficient conditions for the elaboration of these control policies. Here margins are fixed by the amplitude and stability of the domestic financial market; the need for autonomy is defined by the type of policies pursued.

The history of exchange controls emphasizes this double aspect: the impact of current economic policies, and the importance of the financial market with its international links.

One of the first experiences of exchange control in modern times took place in 17th-century England. The Royal Exchange was then introduced, together with the Navigation Acts, in order to secure a basis for growing British power in the face of the decline of Spain and the Netherlands, at a time when the City did not yet carry the weight required for such a rise to power. In Germany in the 1930s, rigorous exchange controls were due to the autarkic tendencies pursued by the national-socialist government. Yet in France, the introduction of such controls after 1936 when the tripartite alliance (with the US and Britain) had failed, calls into question the ability of the financial market to withstand speculative moves such as those following the 1930s crisis.

In the immediate postwar period, generalized controls in Europe revealed the fragility of financial markets in a period when reconstruction absorbed most resources. The progressive and partial liberalization of capital movements (see Einzig 1962) had to rely on massive and conditional Marshall Plan aid and on the regulatory action of institutions such as the IMF, OECD and GATT.

This attempt to insulate a fragile financial market from competition by foreign capital (without the evolution of the rates of exchange correcting this distortion) can explain the relative continuity of restrictions on the movement of capital in France (see Claassen and Wyplosz 1982).

In the growing integration of financial markets there might be seen a stabilizing factor which enables the opening up of relevant economies to the free circulation of capital. The development of new information techniques in the area of communication has largely contributed to the acceleration of this integration of financial markets (initiated in the 1930s by the opening of the first transatlantic communication line): a world market in currency transactions was established in the 1970s; a securities market was in turn established during the 1980s. But the extension of these information networks has also considerably increased the amplitude and scope for short-term speculative movements, thereby increasing global instability in the financial international system. The resulting prospect of international crisis renders unlikely a definite liberalization of monetary movements. If crises can break out more rapidly than in the past, the possibility of introducing rapid and efficient exchange control can play an important role in deterring speculation and the development of a major exchange crisis.

See Also

- ▶ Capital Flight
- ▶ External Debt
- ▶ Fixed Exchange Rates
- ▶ Fundamental Disequilibrium
- ▶ International Capital Flows
- ▶ International Finance

Bibliography

Aglietta M. 1984. Les régimes monétaires de crise. *Critiques de l'Economie Politique* 26–7, January–June.

Bergsten, F., and J. Williamson. 1983. Exchange rates and trade policy. In Cline (1983).

Bhagwati, J. 1978. *Foreign trade regimes and economic development: Anatomy and consequences of exchange control regimes*, vol. XI. Cambridge, MA: Ballinger for the National Bureau of Economic Research.

Claassen, E.M., and C. Wyplosz. 1982. Capital controls: Some principles and the French experience. *Annales de l'Insée* 47–8.

Clarimonte, P., and J.H. Cavanagh. 1984. Transnational corporations and services: The final frontier. In UNCTAD (1984)

Cline, W.R. (ed.). 1983. *Trade policy in the 1980s*. Cambridge, MA: MIT Press.

De Brunhoff, S. 1986. *L'heure du marché*. Paris: Presses Universitaires de France.

Einzig, P. 1934. *Exchange control*. London: Macmillan.

Einzig, P. 1962. *The history of foreign exchange*. London: Macmillan. Reprinted 1979.

Einzig, P. 1968. *Leads and lags*. London: Macmillan.

Johnson, H.G. 1969. The case for flexible exchange rates. *Federal Reserve Bank of St Louis Review* 51: 12–24.

Kindleberger, C.P. 1984. *A financial history of Western Europe*. London: George Allen & Unwin.

Krueger, A.O. 1978. *Liberalization attempts and consequences. vol X of Foreign trade regimes and economic development*. Cambridge, MA: Ballinger for the Nationae Bureau of Economic Research.

McKinnon, R. 1963. Optimum currency areas. *American Economic Review* 53(September): 717–725.

McKinnon, R. 1979. Foreign trade regimes and economic development: A review article. *Journal of International Economics* 9(3): 429–452.

Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51: 657–665.
Mundell, R.A. 1968. *International economics*. New York: Macmillan.
Parboni, R. 1981. *The dollar and its rivals*. London: Verso.
Robinson, J. 1937. The foreign exchanges. In *Essays in the theory of employment*, ed. J. Robinson. London: Macmillan.
Tobin, J. 1978. A proposal for international monetary reform. Cowles Foundation discussion paper no. 506, Yale University.
UNCTAD. 1984. *Trade and development*. Geneva: An Unctad Review.

Exchange Market Pressure

Henk Jager and Franc Klaassen

Abstract

Currencies can be under severe pressure in the foreign exchange market, but in a fixed (or managed) exchange rate regime that is not fully visible via the change in the exchange rate. Exchange market pressure (EMP) is a concept developed to nevertheless measure the pressure in such cases. This article describes EMP and its measurement.

Keywords

Central bank; Currency crisis; Exchange rate regime; Interest rate; Intervention; Monetary policy

JEL Classifications

E52; E58; F31; F33

Definition and Relevance

Exchange market pressure (EMP) on a currency is its excess supply in the foreign exchange market if monetary authorities did not try to influence the exchange rate; this excess supply is expressed in the relative depreciation required to remove it.

Under a floating exchange rate the monetary authorities (usually the central bank) are indeed passive to the exchange rate, so EMP is the actual depreciation. In any other regime the monetary authorities ward off depreciation by policy measures, such as setting a higher official interest rate, or buying domestic currency in the foreign exchange (forex) market. Then the actual depreciation does not coincide with EMP, and correct EMP measurement requires adding the depreciation-counteracting policy actions. The question in the EMP literature, originating from Girton and Roper (1977), is how to do that.

Focusing on EMP rather than sheer exchange rate changes is practically relevant, as 82 per cent of the world's currencies have some sort of peg or managed float (IMF 2009). A first application of EMP exploits the fact that EMP covers the whole spectrum of exchange rate regimes, from floating to fixed. As exchange rate and balance of payment theories essentially focus on tensions in the forex market, under either floating or fixed rates, EMP can integrate both types of theory. Second, EMP can be more relevant than exchange rate changes as a determinant of other phenomena. For instance, IMF (2007) takes EMP to study adequate policy responses to surges in capital inflows. Since EMP better signals forex tensions than exchange rate changes, EMP also helps speculators to find profit opportunities, and policy makers to take timely moves to counteract contagion from other countries.

Measure

A crucial element in the EMP definition is that EMP is a counterfactual concept. That is, it is not the actual situation, where the central bank may ward off pressure, that matters, but the hypothetical situation where the central bank (unexpectedly) does not try to influence the exchange rate, as stressed in Weymark (1995). This makes EMP unobservable (except for a pure float). However, we do observe the policy responses to pressure, besides the exchange rate change. This provides an opportunity to quantify EMP in an indirect way.

One typically includes three pressure-offsetting variables, namely the exchange rate, interest rate, and official forex intervention, though some authors exclude the interest rate. Let s_t denote the (logarithm of the) nominal spot exchange rate at time t, defined as the domestic currency price of one unit of foreign currency. The interest rate i_t is supposed to summarize the use of all money market instruments by the central bank, so it is typically a short-term rate. Finally, c_t is the central bank purchase of domestic currency in the forex market, usually approximated by the decrease in official reserves scaled by a proxy of forex market turnover. This all concerns policy of the domestic central bank. For simplicity, the foreign central bank is assumed not to try to affect the exchange rate.

The pressure-offsetting variables lead to EMP measure

$$EMP_t = \Delta s_t + w_i \tilde{i}_t + w_c \tilde{c}_t,$$

where Δ is the first-difference operator, Δs_t, $\tilde{i}_t$, and $\tilde{c}_t$ are the EMP components based on s_t, i_t and c_t, to be specified below, and w_i and w_c are the EMP weights. This measure does not depend on the sources of pressure, nor is a model of exchange rate determination needed to derive it, as Klaassen and Jager (2008) show using just a few assumptions.

Components

The presence of Δs_t is logical given the EMP definition. It has weight unity, so that indeed the EMP measure is in units of depreciation and coincides with the actual depreciation in case of a floating exchange rate regime. A zero counterfactual official forex intervention implies $\tilde{c}_t = c_t$.

The interest rate component $\tilde{i}_t$ differs across studies. The traditional choice is $\tilde{i}_t = \Delta i_t = i_t - i_{t-1}$, which can essentially be traced back to Girton and Roper (1977). It implies that during a speculative attack where the interest rate is set at, say, 100 per cent for two consecutive days, $\Delta i_t = 0$ on day two, so this EMP component would suggest there is no pressure on that day.

The underlying reason for this counterintuitive result is that in the counterfactual, as prescribed by the EMP definition, the interest rate is not i_{t-1}. The true counterfactual rate is the one the central bank would choose to achieve other targets than the exchange rate, usually domestic targets, such as inflation and output. Therefore, Klaassen and Jager (2008) introduce i_t^d as the counterfactual interest rate and $\tilde{i}_t = i_t - i_t^d$ as a component to obtain an EMP measure that is consistent with the EMP definition. A natural proxy for i_t^d is a Taylor-type rule, but in practice simply taking the foreign interest rate (possibly adjusted by the inflation differential) can be a satisfying approximation.

Weights

The weights w_i and w_c in the EMP measure above state how effective the components are in taking away pressure. The weights are assumed to be positive, but they are not observed.

One way to quantify the weights is by a structural economic model in the spirit of Girton and Roper (1977) and Weymark (1995). A popular choice is a model based on the monetary model of exchange rate determination. The advantage is that the weights have a clear economic meaning, which is useful to the extent that the specification of the model is correct.

Another popular approach is the volatility-smoothing method due to Eichengreen et al. (1996). Here a weight is estimated by the ratio of the sample standard deviation of Δs_t to that of the component involved, so that no component dominates the others in terms of volatility. These weights no longer depend on a structural model and are easier to compute, though they now reflect not only the effectiveness of the monetary policy instruments – as they should – but also how intensively the instruments are used.

See Also

- ▶ Capital Controls
- ▶ Currency Boards
- ▶ Currency Crises

▶ Exchange Rate Target Zones
▶ Nominal Exchange Rates

Bibliography

Eichengreen, B., A.K. Rose, and C. Wyplosz. 1996. Speculative attacks on pegged exchange rates: An empirical exploration with special reference to the European Monetary System. In *The new transatlantic economy*, ed. M.-B. Canzoneri, W.J. Ethier, and V. Grilli, 191–228. Cambridge: Cambridge University Press.

Girton, L., and D. Roper. 1977. A monetary model of exchange market pressure applied to the postwar Canadian experience. *American Economic Review* 76: 537–548.

International Monetary Fund (IMF). 2007. *World economic outlook*. Washington, DC: IMF.

IMF. 2009. De facto classification of exchange rate regimes and monetary policy frameworks. Available at < http://www.imf.org/external/np/mfd/er/2008/eng/0408.htm. Accessed 28 Jan 2010.

Klaassen, F., and H. Jager. 2008. Definition-consistent measurement of exchange market pressure. Available at < http://www.feb.uva.nl/pp/klaassen/. Accessed 28 Jan 2010.

Weymark, D.N. 1995. Estimating exchange market pressure and the degree of exchange market intervention for *Canada. Journal of International Economics* 39: 273–295.

Exchange Rate Dynamics

Nelson C. Mark

Abstract

Exchange-rate dynamics refers to the response path of the exchange rate following the revelation of some economic shock when the country in question operates under a pure flexible exchange-rate system. The issue attracts research attention because of the volatile nature of the exchange rate and the belief that the exchange rate may affect the allocation of resources across countries and over time. If observed exchange-rate dynamics cannot be shown to have a rational basis, efficiency will suffer from decisions made conditioned on disequilibrium values of the exchange rate.

Keywords

Bretton Woods system; Dornbusch, R.; Exchange rate overshooting; Exchange rate volatility; Exchange-rate dynamics; Fixed exchange rates; Flexible exchange rates; Incomplete markets; Local-currency pricing; Monetary shocks; Mundell–Fleming model; New open-economy macroeconomics; Producer-currency pricing; Real exchange rates; Redux model; Risk aversion; Sticky prices; Uncertainty; Uncovered interest parity; Vector autoregressions

JEL Classifications

F31; F4

Exchange-rate dynamics refers to the response path of the exchange rate following the revelation of some economic shock (news) when the country in question operates under a pure flexible exchange-rate system.

The topic has attracted research attention since 1973 when the industrialized world abandoned the Bretton Woods system of fixed exchange rates. The initial experience in the 1970s with this new exchange-rate system surprised economists along two dimensions. The first surprise was that exchange-rate returns turned out to be much more volatile than expected. This volatility, which is similar in magnitude to the volatility of stock returns, is much higher than the volatility of macroeconomic fundamentals such as the growth rate of money or income. The second surprise was the very high persistence of the exchange rate. The logarithm of the exchange rate evolves approximately as a random walk so that all shocks appear to have a permanent effect on the exchange rate level. As a result, percentage changes of the exchange rate (returns) over short horizons are nearly unpredictable. These features of quarterly data from 1973Q1 to 2002Q1 are illustrated in Table 1.

Academic and policy interest in understanding exchange-rate dynamics stems from the belief that the exchange rate affects the current account, a country's international indebtedness, and the rate of capital formation, and is therefore an important

Exchange Rate Dynamics, Table 1 Volatility (sample standard deviation) and autocorrelations of selected currencies, and other variables, 1973–2002

	DM	£	¥	S&P	M1	US GDP	T-Bill
Volatility	23.384	25.040	20.660	31.719	12.200	3.384	3.118
ρ_1	0.156	0.145	0.167	0.113	− 0.273	− 0.447	0.539
ρ_2	− 0.072	− 0.034	− 0.128	− 0.026	0.301	0.061	0.473
ρ_2	0.126	0.149	0.082	0.068	− 0.309	− 0.084	0.534
ρ_4	0.118	0.057	0.030	0.071	0.741	0.004	0.595

Notes: Data are annualized quarterly growth rates. They show volatility (sample standard deviation) and autocorrelations of real annualized returns on the Deutschmark, yen, and pound sterling relative to the US dollar, returns on the Standard and Poor's index, real M1 growth, and real US gross domestic product growth, and the three-month Treasury-bill rate. All variables are in real terms

Sources: Standard and Poor's from Robert Shiller's website http://www.econ.yale.edu/Bshiller/data.htm. All other data from Federal Reserve Economic Data (FRED)

macroeconomic variable that has real allocative implications for the open economy. An important question is whether observed exchange-rate dynamics have a rational basis or if it reflects an irrational overreaction to shocks. If it is the latter, economic efficiency may be compromised when allocative decisions are made conditional on disequilibrium or non-fundamental values of the exchange rate.

Due to limited experience with flexible exchange rates combined with high degrees of capital mobility during the Bretton Woods era, the volatile nature of exchange rates was difficult to anticipate. Moreover, the generally accepted framework at the time, for modeling the open economy, was the static Mundell–Fleming model which was poorly equipped for understanding these issues. The immediate post Bretton Woods experience stimulated a large body of theoretical work aimed at improved modelling the exchange rate. This work culminated with Dornbusch's (1976) celebrated exchange-rate overshooting model which provided a rational explanation for high exchange-rate volatility in the presence of relatively stable macroeconomic fundamentals. The overshooting model is a deterministic perfect-foresight dynamic generalization of Mundell–Fleming. The critical features are the differential speeds of adjustment between the goods (gradual) and asset (immediate) markets, uncovered interest parity, and the central importance of monetary shocks as the underlying source of uncertainty. Subsequent econometric work and quantitative analyses of dynamic general equilibrium models known as the 'new open-economy macroeconomics' suggest that an exact understanding of the mechanism that generates overshooting and excess exchange-rate volatility remains elusive.

Exchange-Rate Overshooting

Let e denote the natural logarithm of the home-currency price of the foreign currency. Under this definition of the nominal exchange rate, an increase in e means the home currency has weakened. Let the forward-looking steady state log exchange rate be $\bar{e}$. Assume that commodity prices are sticky, output is fixed, real money demand is inversely related to the interest rate i, financial capital is perfectly

internationally mobile and domestic and foreign currency assets are perfect substitutes. In a deterministic setting, the perfect foresight instantaneous change in the exchange rate e can be shown to be proportional to the current exchange-rate gap $(\bar{e} - e)$. Let the exogenous foreign interest rate be i^*. Then uncovered interest parity, which says that an excess yield on domestic nominal assets is offset by an expected capital loss on the domestic currency through movements in the exchange rate, can be expressed as

$$i - i^* = \theta(\bar{e} - e).$$

This is the asset market equilibrium condition.

Now consider a one-time permanent surprise increase of one per cent in the home country's

money supply. At the instant the shock is revealed, $\overline{e}$ increases by 0.01 in accordance with long-run monetary neutrality. Noting that the price-level is instantaneously fixed, the monetary shock creates a liquidity effect that lowers the interest rate i. To maintain uncovered interest parity, the instantaneous exchange rate e must increase by more than the 0.01 increase in $\overline{e}$, thereby 'overshooting' its long-run steady state value. Thus, the short-run variability of the exchange rate is seen to exceed the variability of the macroeconomic fundamentals in response to monetary shocks. In order to provide a general explanation for exchange rate volatility, however, it must be the case that nominal shocks are relatively important drivers of aggregate uncertainty because the model does not generate overshooting in response to real shocks such as shifts in the aggregate expenditure function. The overshooting model represented a significant contribution by giving an explanation for high exchange-rate variability in the context of a rational equilibrium model.

In stochastic generalizations of this model, as considered by Mussa (1982) and Obstfeld (1985), one would say that the exchange rate exhibits excess volatility in the sense that the volatility of the exchange rate exceeds that of the macroeconomic fundamentals (money supply and income). Moreover, in the stochastic environment the equilibrium exchange rate has a present-value representation that discounts the expected future flow of the macroeconomic fundamentals. The analogy to the present-value model of stock and bond pricing rationalizes why the exchange rate and other asset prices have many common properties.

Emerging Doubts About the Overshooting Mechanism

Doubts about the precise mechanism that generates overshooting have been raised in vector autoregression (VAR) studies. Eichenbaum and Evans (1995) modelled monetary shocks both as innovations to non-borrowed reserves and as innovations in the federal funds rate, and used a recursive ordering strategy to identify the VAR. Following a monetary shock, their impulse–response analysis did now show the immediate overshooting and the subsequent exponential decay of the exchange rate predicted by the overshooting model. Instead, they found a 'hump-shaped' response in the exchange rate: while it did overshoot the long-run equilibrium value, it did so gradually. This so-called 'delayed overshooting' result was confirmed in subsequent studies with structural VARs, such as Clarida and Gali (1994).

While such VAR studies create a dilemma for the overshooting mechanism, there is the possibility that the appearance of delayed overshooting was created by imposing false restrictions in the identification of VARs. Faust and Rogers (2003) argue that restrictions implied by recursive orderings employed in VAR analyses are implausible because they rule out contemporaneous responses of many important variables, such as the foreign interest rate, to domestic monetary policy shocks. Experimenting with alternative sets of more plausible restrictions on contemporaneous interactions, they find that immediate overshooting sometimes occurs and sometimes does not. In those situations where immediate overshooting does occur, they find that it is driven by deviations in uncovered interest parity. The implication then is that, if immediate overshooting is actually present in the data, it does not appear to be generated by the mechanism articulated by Dornbusch.

The preceding development of overshooting was discussed in terms of the nominal exchange rate, but the qualitative predictions would not be affected if it were recast in terms of the real exchange rate. At this point, we see that three pillars of the overshooting explanation of exchange-rate volatility are sticky nominal goods prices, uncovered interest parity, and the principal importance of monetary shocks. The question of whether prices are sticky is a tough problem that macroeconomists struggle to answer. Some international macroeconomic evidence has emerged to suggest that sticky goods prices are an important feature of the data. It has been pointed out by Mussa (1986) that the real exchange rate eP^*/P is much more volatile under flexible exchange rates than under fixed exchange rates. When the exchange rate is flexible, the correlation between nominal and real exchange rate movements is

exceedingly high over relatively short horizons, say, from a month to a year (except in very high-inflation countries). It has also been observed that international violations of the law of one price are more severe than within-country violations. Engel and Rogers (1996) find that the volatility of the difference between the log price of a particular good sampled in two locations is much higher if the sample points are in different countries than if they are within the same country.

Exchange-Rate Dynamics in 'New Open-Economy Macroeconomic' Models

The exchange-rate models of the 1970s and 1980s were typically built from sets of ad hoc macroeconomic relations that lack rigorous microeconomic foundations. As exchange-rate models have evolved towards dynamic general equilibrium analysis, the research continues to try to understand the necessary conditions for an equilibrium rational-expectations monetary model to explain observed exchange-rate volatility and persistence. A major vein of this work is done within the 'new open-economy macroeconomics' (NOEM) class of theories that features rational optimizing agents, imperfect competition and sticky goods prices in a dynamic general equilibrium open-economy setting.

The NOEM began with the Obstfeld and Rogoff (1995) Redux model. Money is introduced through the utility function, and agents maximize expected lifetime utility defined over consumption, leisure and real money balances. Goods prices are set one period in advance by monopolistically competitive firms that produce output with labour but no physical capital. The precise way in which the exporting firm sets the price of output for foreign and domestic buyers has significant implications for exchange-rate dynamics.

In the Redux model, firms engage in 'producer-currency pricing'. That is, an exporter will set the domestic price of the good at P and foreigners pay P/e units of the foreign currency. While P is fixed for the period, e is not, so a within-period change in the nominal exchange rate will alter the relative price between home and foreign goods. This is the basis of the 'expenditure-switching' effect of the exchange rate stressed in the Mundell–Fleming model. Even though goods prices are sticky, the Redux model generates the surprising result that a monetary shock causes the nominal exchange rate to jump immediately to its long-run value. In other words, there is no overshooting or excess volatility of the exchange rate in the Redux model.

An alternative to producer-currency pricing is 'pricing-to-market' (also known as 'local-currency pricing'). In this scenario, segmentation between the domestic and foreign goods markets allows firms to set a foreign currency export price that is different from the domestic currency price of the good. Here, foreigners pay the pre-set price P^* and the exporting firm receives eP^* units of home currency. Within-period changes in the exchange rate affect firm revenues but not the relative price between imported and domestic goods. Betts and Devereux (2000) apply this idea in modifying the Redux model and find that increasing the fraction of firms that price-to-market reduces the strength of the expenditure-switching effect, which attenuates the allocative role of the exchange rate. As a result, a larger change in the exchange rate is required to restore equilibrium following a monetary shock. Accordingly, if the degree of pricing-to-market is sufficiently high, the overshooting result can be restored.

Some NOEM modellers ask their models to quantitatively match the data. A quantitative analysis of the exchange-rate dynamics implied by an NOEM model might suppose that agents operate in a complete markets setting where they can trade a full set of state-contingent claims, which allows complete international risk-sharing in a model. Production requires labour and durable physical capital. There is a final good that is not traded internationally and produced in a competitive industry. Intermediate goods are traded internationally and produced by monopolistically competitive price-setting firms that change prices according to a staggered price-setting rule. The period utility is defined over consumption C_t, real money balances M_t/P_t, and leisure L_t. With asterisks denoting foreign country variables, the

exchange rate is determined by the equilibrium risk-sharing condition

$$\frac{E_t P_t^*}{P_t} = \Gamma \frac{U_{c^*}\left(C_t^*, M_t^*/P_t^*, L_t^*\right)}{U_c(C_t, M_t/P_t, L_t)}.$$

where E_t is the nominal exchange rate level and U_c is the marginal utility of consumption. The real exchange rate is proportional to the ratio of foreign to domestic marginal utility. If utility is separable in its arguments and has the constant relative risk-aversion form, the real exchange rate will be proportional to relative consumption levels in the two countries and the factor of proportionality Γ will be increasing in the coefficient of relative risk aversion. This exchange-rate pricing condition is invariant to whether commodity prices are sticky or flexible. The question is under what conditions will the model generate consumption responses to monetary shocks to create exchange-rate dynamics like those found in the data.

Chari, Kehoe and McGrattan (2002) calibrate this two-country model to the United States and an aggregate 'European' country which they employ to simulate model-implied observations of the endogenous variables. Using the staggered-price setting rule of Taylor (1999), they find that firms cannot be allowed to change prices more frequently than once a year and that the coefficient of relative risk aversion must be at least 5 in order for the model-generated real exchange rate to match the persistence and volatility found in the data. While these are not unreasonable conditions, the degrees of price stickiness and of risk aversion required both seem to be on the high side. Kollman (2001), on the other hand, is able to obtain exchange-rate overshooting with a relative risk-aversion coefficient of 2, but uses the Calvo (1983) rule for price setting.

While this line of work shows that the interaction of monetary shocks and sticky prices is able to explain volatile and persistent exchange-rate behaviour, the models often generate counterfactual predictions for other dimensions of the data. One of these counterfactuals, obtained from the exchange-rate determination equation, is that the real exchange rate should be nearly perfectly correlated with relative foreign-to-home consumption levels. In the data, however, the real exchange rate and relative consumption levels are uncorrelated – a problem known as the Backus and Smith (1993) puzzle.

The restrictive feature of models with complete markets is the international risk-sharing condition which constrains the extent to which the exchange rate can move. One might think that limiting the menu of tradable assets and working in an incomplete markets environment might loosen up this restriction. A common specification of an incomplete markets environment is to allow only a non-state-contingent bond to be internationally traded. However, as demonstrated by Baxter and Crucini (1995), the quantitative properties of dynamic general equilibrium models are little affected when full risk sharing is replaced by this version of incomplete markets, especially when shocks to the environment are not permanent. The reason is that uncovered interest parity replaces the risk-sharing condition in determining the exchange rate, so relatively smooth interest rates replace smooth consumption levels in limiting the range of exchange-rate movements.

Accordingly, studies by Devereux and Engel (2002), Kollman (2001) and Duarte and Stockman (2005) of exchange-rate dynamics in dynamic general equilibrium models under incomplete markets find it necessary to allow exogenous deviations from uncovered interest parity in order to explain exchange-rate volatility in the presence of smooth fundamentals. There is ample empirical work, surveyed by Hodrick (1987), Engel (1996), and Lewis (1995), to justify these violations, and theoretical work by Mark and Wu (1998) and Jeanne and Rose (2002) gives an explanation for the deviations through participation of noise traders in the foreign-exchange market.

High exchange-rate volatility, persistence, and overshooting can be generated from fully specified dynamic stochastic general equilibrium models. However, a fully convincing story about exchange-rate dynamics remains elusive because an accepted model that provides reasonably good ability to account for all of the salient features of the data is not yet available.

See Also

- ▶ Exchange Rate Volatility
- ▶ Foreign Exchange Market Microstructure
- ▶ Real Exchange Rates

Bibliography

Backus, D., and G. Smith. 1993. Consumption and real exchange rates in dynamic economies with non-traded goods. *Journal of International Economics* 35: 297–316.

Baxter, M., and M. Crucini. 1995. Business cycles and the asset structure of foreign trade. *International Economic Review* 36: 821–854.

Betts, C., and M. Devereux. 2000. Exchange rate dynamics in a model of pricing-to-market. *Journal of International Economics* 50: 215–244.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Chari, V., P. Kehoe, and E. McGrattan. 2002. Can sticky price models generate volatile and persistent real exchange rates? *Review of Economic Studies* 69: 533–564.

Clarida, R., and J. Gali. 1994. Sources of real exchange-rate fluctuations: How important are nominal shocks? *Carnegie-Rochester Conference Series on Public Policy* 41: 1–56.

Devereux, M., and C. Engel. 2002. Exchange rate pass-through, exchange rate volatility, and exchange rate disconnect. *Journal of Monetary Economics* 49: 913–940.

Dornbusch, R. 1976. Expectations and exchange-rate dynamics. *Journal of Political Economy* 84: 1161–1176.

Duarte, M., and A. Stockman. 2005. Rational speculation and exchange rates. *Journal of Monetary Economics* 52: 3–29.

Eichenbaum, M., and C. Evans. 1995. Some empirical evidence on the effects of shocks to monetary policy on exchange rates. *Quarterly Journal of Economics* 110: 975–1009.

Engel, C. 1996. The forward discount anomaly and the risk premium: A survey of recent evidence. *Journal of Empirical Finance* 3: 123–192.

Engel, C., and J. Rogers. 1996. How wide is the border? *American Economic Review* 86: 1112–1125.

Faust, J., and J. Rogers. 2003. Monetary policy's role in exchange rate behavior. *Journal of Monetary Economics* 50: 1403–1424.

Hodrick, R. 1987. *The empirical evidence on the efficiency of forward and futures foreign exchange markets*. Chur: Harwood Academic Publishers.

Jeanne, O., and A. Rose. 2002. Noise trading and exchange rate regimes. *Quarterly Journal of Economics* 117: 537–569.

Kollman, R. 2001. The exchange rate in a dynamic-optimizing business cycle model with nominal rigidities: A quantitative investigation. *Journal of International Economics* 55: 243–262.

Kollman, R. 2005. Macroeconomic effects of nominal exchange rate regimes: New insights into the role of price dynamics. *Journal of International Money and Finance* 24: 275–292.

Lewis, K. 1995. Puzzles in international financial markets. In *Handbook of international economics*, ed. G. Grossman and K. Rogoff, Vol. 3. Amsterdam: North-Holland.

Mark, N., and Y. Wu. 1998. Rethinking deviations from uncovered interest parity: The role of covariance risk and noise. *Economic Journal* 108: 1686–1706.

Mussa, M. 1982. A model of exchange rate dynamics. *Journal of Political Economy* 90: 74–104.

Mussa, M. 1986. Nominal exchange rate regimes and the behavior of real exchange rates: Evidence and implications. *Carnegie Rochester Conference Series on Public Policy* 25: 117–213.

Obstfeld, M. 1985. Floating exchange rates: Experience and prospects. *Brookings Papers on Economic Activity* 1985(2): 369–450.

Obstfeld, M., and K. Rogoff. 1995. Exchange rate dynamics Redux. *Journal of Political Economy* 103: 624–660.

Taylor, J. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford, Vol. 1B. Amsterdam: North-Holland.

Exchange Rate Exposure

Kathryn M. E. Dominguez and Linda L. Tesar

Abstract

Exchange rate exposure describes the influence of exchange rate movements on the value of a firm or sector of the economy. Exposure is typically measured as the correlation of firm or industry stock returns and exchange rate changes in the context of a market model. Exposure appears to be most prevalent in firms that are small (these are less likely to engage in hedging activities) or involved in international activities. While studies have linked *ex ante* exchange rate risk with firm investment strategies, it has proven difficult to

identify the *ex post* consequences of exposure on firm or industry behaviour.

Keywords

Capital asset pricing model; Exchange rate exposure; Exchange rate risk; Hedging; Operating exposure; Trade-weighted basket of currencies; Transaction exposure; Translations exposure; Value-weighted market returns

JEL Classifications

F3

Exchange rate exposure measures the extent to which firm value is influenced by exchange rate movements. Narrow definitions of exchange rate exposure, such as transaction exposure, translations exposure, or operating exposure, focus on the effect of the exchange rate on specific components of a firm's balance sheet or on different types of transactions. More broadly, economists use the term 'economic exposure' to describe the impact of exchange rate movements on a firm or a sector of the economy. A firm's 'exposure' could be measured as the responsiveness of profits, cash flow, operating costs, total assets or liabilities, markups or even wage-setting behaviour to fluctuations in the exchange rate.

If one has detailed information about a firm's operations, it may be possible to trace the effect of exchange rate shifts on a specific line item in the firm's accounting data, with other factors controlled for. More generally, however, measures of exchange rate exposure involve an indirect test of the link between the exchange rate and a firm's stock return, where the return is taken as a measure of the firm's profitability. Under the capital asset pricing model (CAPM), the expected risk premium on a company's share price is proportional to its covariance with the market portfolio. In theory, investors will require a return only on the non-diversifiable portion of firm risk and, if the market return captures all systematic risk, no variable other than the market return should play a role in determining asset returns. Therefore, a test for exchange rate exposure involves including the change in the exchange rate on the right-hand side of a standard CAPM regression and testing whether its coefficient is significantly different than zero (Adler and Dumas 1984):

$$R_{i,t} = \beta_{0,i} + \beta_{1,i} R_{m,t} + \beta_{2,i} \Delta s_t + \varepsilon_{i,t} \qquad (1)$$

where $R_{i,t}$ is the return on firm i at time t, $R_{m,t}$ is the return on the market portfolio, $\beta_{1,i}$ is the firm's beta, Δs_t is the change in the relevant exchange rate and $\beta_{2,i}$ measures a firm's exposure to exchange rate movements after the overall market's exposure to currency fluctuations is taken into account. If $\beta_{2,i}$ is zero, this implies that firm i has the same exchange rate exposure as the market portfolio (not necessarily that the firm has no exposure). Alternatively, if one rejects the hypothesis that $\beta_{2,i}$ is, on average, zero, this indicates that the firm is more exposed to exchange rate movements than the market. Note that the interpretation of $\beta_{2,i}$ as a measure of exposure is conditional on Eq. 1 being the 'true' model of asset returns. Evidence that $\beta_{2,i}$ is non-zero can be interpreted as evidence against the joint hypothesis that the CAPM holds (that is, the market efficiently prices systematic risk) and that exchange rate risk is unimportant for stock returns.

An alternative approach is to measure total exposure, or the unconditional correlation of exchange rates and returns that would involve omitting the market return as an explanatory variable in Eq. 1. The advantages to measuring total exposure are that it allows one to measure the exposure of all firms as a group rather than individual firms relative to the group average, and it requires no assumption about the validity of the CAPM. The disadvantage of total exposure is that it does not allow one to distinguish between the direct effect of exchange rate changes and the effects of macroeconomic shocks that simultaneously affect firm value and exchange rates (which we assume are captured in Eq. 1 by the market return).

Predicting which firms are likely to be affected by changes in the exchange rate and the direction that exposure might take can be quite complicated. There are a number of channels through which the exchange rate might affect the profitability of a firm. Firms that export to foreign markets may benefit from a depreciation of the

local currency if their products become more affordable to foreign consumers. On the other hand, firms that rely on imported intermediate products may see their profits shrink as a consequence of increasing costs of production. Firms with foreign subsidiaries may shift activities across national boundaries, taking on exchange rate risk to take advantage of tax differentials. Firms that do no international business may be influenced indirectly by foreign competition. Furthermore, firms in the non-traded as well as the traded sectors of the economy compete for factors of production, whose returns may be affected by changes in the exchange rate.

A number of specific issues arise when implementing Eq. 1 as a test for exchange rate exposure. First, one must identify the relevant exchange rate. Many studies use a trade-weighted exchange rate to measure exposure. The problem with using a trade-weighted basket of currencies in exposure tests is that the results lack power if the nature of firm exposure does not correspond to the exchange rates (and the relative weights) included in the basket. More generally, one should expect variation in individual firm and industry exposure to various exchange rates. Dominguez and Tesar (2001a) find that any test that restricts the measurement of exposure to one exchange rate (whether it be a trade-weighted rate or a bilateral rate) is likely to be biased downward. Exposure may also be a function of horizon. It may be that firms can use financial derivatives to hedge exchange rate risk in the short run, but remain exposed to low-frequency movements in exchange rates over long horizons. Dominguez and Tesar (2006) find that exposure generally increases with return horizon.

A second issue is sample selection – which firms should be included in empirical tests for exposure. Much of the literature has focused on exposure in multinational firms (Jorion 1990; He and Ng 1998), or in firms that actively engage in international trade (Bodnar et al. 2002). However, there are good reasons to think that exposure will not be limited to these firms. Dominguez and Tesar (2001b) find that firms that engage in greater trade exhibit *lower* degrees of exposure, reflecting the fact that they are also the most aware of exchange-rate risk and, therefore, the most likely to hedge their exposure. Exposure may also be affected by the competitive structure at the industry level. Less competitive industries with higher markups can adjust prices in response to exchange rates, and the impact on profitability and returns will thus be smaller. Allyanis and Ihrig (2001) show that the extent of exposure varies systematically with the extent of industry-level markups.

A third issue that arises in tests for exchange rate exposure is the specification of the market index. Empirical tests of the standard CAPM model generally include a country-specific value-weighted market return to proxy for 'the market'. Value-weighted market returns are likely to be dominated by large firms that are more likely to be multinational or export-oriented, and are more likely than other firms to experience negative cash flow reactions to home currency appreciations. In a world of perfectly integrated capital markets the 'market return' should be proxied by a global portfolio. Dominguez and Tesar (2006) sort out the impact of the choice of market index on exposure for a sample of firms in eight (non-US) markets and find little difference in estimated exposure level using value-weighted and equal-weighted indices, but find a significant increase in measured exposure using a global index. Part of the explanation for this is that the global index generally does a poor job of explaining returns, so that more firms appear to be exposed simply because the exchange rate is picking up more of the variability in returns, and the (global) market index is picking up substantially less.

A final issue is whether exposure is predictable. The standard way to test this is to run a second-stage regression that takes the estimated exposure betas from Eq. 1 and regresses these on a variety of potential explanatory variables. Using this approach Dominguez and Tesar (2006) find that exposure is more prevalent in small (rather than large or medium-sized) firms, and in firms engaged in international activities (measured by multinational status, holdings of international assets, and foreign sales).

Once one has identified a set of firms that are exposed to exchange rate risk, the question remains whether such exposure affects firm behaviour in

some way, such as its level of investment or market entry. In general, it has proven difficult to identify a link between *ex post* exposure (as measured by estimates of exposure in Eq. 1) and such economic outcomes. Numerous studies, however, have linked firm strategies for handling exchange rate risk *ex ante* with their investment decisions in domestic and foreign markets.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Hedging
- ▶ Risk

Bibliography

Adler, M., and B. Dumas. 1984. Exposure to currency risk: Definition and measurement. *Financial Management* 13: 41–50.

Allyanis, G., and J. Ihrig. 2001. Exposure and markups. *Review of Financial Studies* 14: 805–835.

Bodnar, G., B. Dumas, and R. Marston. 2002. Pass-through and exposure. *Journal of Finance* 57: 199–231.

Dominguez, K., and L. Tesar. 2001a. A re-examination of exchange rate exposure. *American Economic Review Papers and Proceedings* 91: 396–399.

Dominguez, K., and L. Tesar. 2001b. Trade and exposure. *American Economic Review Papers and Proceedings* 91: 367–370.

Dominguez, K., and L. Tesar. 2006. Exchange rate exposure. *Journal of International Economics* 68: 188–218.

He, J., and L. Ng. 1998. The foreign exchange exposure of Japanese multinational corporations. *Journal of Finance* 53: 733–753.

Jorion, P. 1990. The exchange rate exposure of U.S. multinationals. *Journal of Business* 63: 33–45.

Exchange Rate Target Zones

John Driffill

Abstract

A target zone attempts to limit the movement of an exchange rate, avoiding the pitfalls of both a pegged rate and a freely floating rate. The European Monetary System was the prime example. An elegant model of Paul Krugman demonstrates that in theory a target zone does indeed stabilize an exchange rate. But in practice it has been substantially rejected empirically. Williamson's 'crawling bands' around a 'fundamental equilibrium exchange rate' develop the concept. Target zones survive among candidates for membership of the Eurozone who take part in the Exchange Rate Mechanism mark II.

Keywords

Bretton Woods System; Brownian motion; Capital controls; Economic and Monetary Union (EMU); Euro; European Monetary System; Exchange Rate Mechanism (EU); Exchange rate target zone; Fixed exchange rates; Floating exchange rates; Fundamental equilibrium exchange rate; Inflation differentials; Monetary theory of the exchange rate; Nominal exchange rates; Purchasing power parity; Real exchange rates; Uncovered interest parity; Wiener process

JEL Classifications

F31

An exchange rate target zone is a scheme intended to limit the flexibility of an exchange rate without going as far as fixing or pegging the value of one currency against another. It is a band, or zone, of values for the exchange rate, around a central or target rate. Within the zone, the exchange rate is allowed to fluctuate freely without any intervention from the authorities or, at least, with less intervention than there is elsewhere. At the edge of the band, and outside, if the rate strays there, there is more vigorous intervention to keep the rate within, or return it to, the band. There are many varieties of target zone. The edges may be hard or soft. It may be defined in terms of nominal or real exchange rates. The central rate – the target – may be either constant over time, possibly with provision for occasional discrete changes; or it may be adjusted continuously. The bands may be narrow or wide.

The most celebrated and ambitious target zones were those introduced in 1979 by the European Monetary System (EMS). They operated until the end of 1998. Under the EMS, member states were initially required to keep their bilateral exchange rates within a band of ±2.25 per cent around a grid of central parities (Giavazzi and Giovannini 1989). They were required to use unlimited intervention in the foreign exchange markets to defend the bands if an exchange rate strayed to the edge. Member countries could adjust central parities occasionally by mutual agreement when perceived misalignments had built up. The system evolved over time. As capital controls were progressively removed, orderly realignments became more difficult to manage. The system became less flexible, notably after 1987. The gradual movement towards complete fixity of exchange rates, intended to prevail under Economic and Monetary Union, was thrown off course by massive speculative attacks in September 1992 and August 1993. The system was unable to withstand them and the bands were widened to 15 per cent. But they were subsequently narrowed again and the EMS gave way to the euro on the 1 January 1999.

The use of target zones sprang from a desire to avoid the pitfalls of fixed rates and free floating. Under the fixed exchange rates of the Bretton Woods System (1944–73), exchange rate misalignments had become progressively worse as inflation rates diverged, and weak currency countries put off devaluation, deterred by costly speculation. Under floating exchange rates during the 1970s, exchange rates fluctuated excessively, unrelated to fundamentals like relative price levels and current accounts. The 'disconnect' between exchange rates and economic fundamentals has been confirmed by widespread experience and has become a central tenet of international macroeconomics.

The EMS was intended to allow exchange rates to offset inflation differentials among members. Realignments were to be sufficiently timely to avoid giving the markets a one-way bet. The bands were intended to enable markets to determine exchange rate movements without official intervention for most of the time, at the same as discouraging destabilizing speculation.

The questions of how target zones might work in theory, whether they worked in practice as the theory predicted, and whether they did indeed cut exchange fluctuations, have generated enormous amounts of research.

The key theoretical contribution is that of Krugman (1991). He showed that a fully credible target zone would reduce the volatility of an exchange rate and reduce its sensitivity to fundamentals. His theoretical model assumes a monetary theory of the exchange rate for a small open economy in a world of perfectly flexible prices and perfect capital mobility, in which purchasing power parity and uncovered interest parity hold good. Then the log of the exchange rate (e) can be expressed as a function its own anticipated rate of change over time ($E_t(de)/dt$) and a driving fundamental (f)

$$e = f + \alpha E_t(de)/dt$$

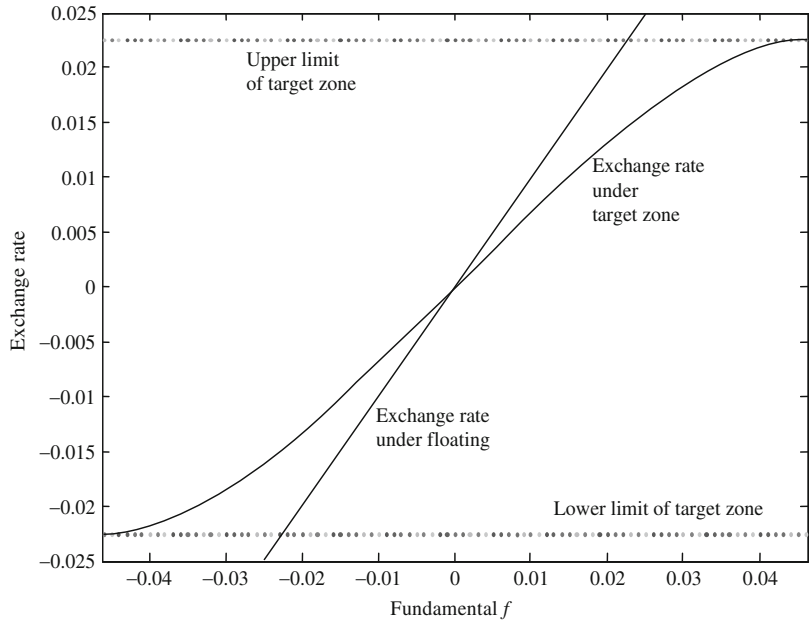
The parameter α denotes the semi-elasticity of money demand with respect to the interest rate. The fundamental reflects money supply and demand. He considers a stochastic model in continuous time, in which the fundamental (f) follows Brownian motion, the continuous-time analogue of a random walk. That is

$$df = \sigma dz$$

where dz is the innovation in a standard Wiener process, and σ is the variance of the innovation in the fundamental per unit time. dz has mean zero and variance $E(dz^2) = dt$. When the exchange rate is allowed to float freely, the exchange rate will be a linear function of the fundamental

$$e = f.$$

But, when a target zone limits movements of the exchange rate, this solution does not apply. Assume for simplicity that under the target zone the central parity for the logarithm of the exchange rate is equal to zero and the limits of the zone are $\bar{e}$ and $\underline{e}$. Further assume that the zone is symmetrical and $\bar{e} = -\underline{e}$. Using stochastic calculus and methods widely used in the theory of



Exchange Rate Target Zones, Fig. 1 *Source*: Krugman (1991)

options pricing, Krugman shows that the exchange rate is related to the driving fundamental by the relationship

$$e = f + A(\exp(\xi f) - \exp(-\xi f))$$

where

$$\xi = \sqrt{\frac{2}{\sigma^2 \alpha}}.$$

The constant A is such that at the top and bottom of the band the value-matching conditions

$$\overline{e} = \overline{f} + A\left(\exp\left(\xi \overline{f}\right) - \exp\left(-\xi \overline{f}\right)\right)$$

and

$$\underline{e} = \underline{f} + A\left(\exp\left(\xi \underline{f}\right) - \exp\left(-\xi \underline{f}\right)\right)$$

hold, and the smooth pasting conditions also hold. These require that the derivative of the exchange rate with respect to the fundamental at the edges of the band is equal to zero. Viz.

$$1 + A\left(\xi \exp\left(\xi \overline{f}\right) + \xi \exp\left(-\xi \overline{f}\right)\right) = 0$$

and

$$1 + A\left(\xi \exp\left(\xi \underline{f}\right) + \xi \exp\left(-\xi \underline{f}\right)\right) = 0.$$

From these conditions the value of the constant A, and the value of the fundamental at the limits of the band ($\overline{f} = -\underline{f}$), can be determined. The value of the parameter A that emerges from this analysis is negative. Thus the value of the exchange rate, corresponding to any particular value of the fundamental, is closer to the parity rate under a target zone than it would be under a free float.

Krugman's analysis establishes that the exchange rate within the target zone enjoys the 'bias in the band' or the 'honeymoon effect'. The relationship between the fundamental and the exchange rate in the target zone is an S-shaped curve. It is flatter everywhere than the relationship under a free float, which is a 45 degree line. The perfectly credible commitment of the authorities to intervene, should the exchange rate ever reach the edge of the band, so as to prevent any movement beyond it, discourages deviations from the central parity even without any actual intervention. It is illustrated in Fig. 1.

This elegant theory has very strong empirical predictions. Three predictions that do not rely on any assumptions about the value of the unobserved fundamentals are as follows. First, the exchange rate should spend a lot of time near the edges of the band, and relatively little near the centre. The unconditional distribution of the fundamental within the band is uniform, and the distribution of the exchange rate is U-shaped. Second, the uncovered interest parity condition implies that, when the exchange rate is strong (e is close to $\underline{e}$) and thus likely to weaken, the domestic interest rate should exceed the foreign rate, while when the exchange rate is weak, the domestic interest rate should be relatively low. And third, the converse of the second prediction, at any point in time the expected future exchange rate implied by the uncovered interest parity condition

$$E_t(e_{t+dt}) = e_t + \left(r_t - r_t^*\right)dt$$

(r_t is the domestic instantaneous nominal interest rate and r_t^* the foreign one) should lie within the band.

Unfortunately, all three of these predictions are comprehensively rejected by empirical evidence, of which a great deal has been accumulated. The work of Flood et al. (1991) led the way in this. They and many subsequent studies have found that exchange rates had tended to be concentrated near the middle of the band, not at the edges, as predicted. They also found that, when the exchange rate was weak, there was no tendency for the interest rate to be relatively low. The expected future exchange rate implied by uncovered interest parity was found to spend a great deal of time outside the band, suggesting a lack of credibility of the target zone.

Direct tests of the relationship between the exchange rate and the fundamental driving variable have generally found very little evidence of non-linearity or Sshapedness. There appears to be little evidence of any 'honeymoon effect'. Svensson (1992) remarks that the comprehensive rejection of this theory looks like what T.H. Huxley called 'the great tragedy of science – the slaying of a beautiful hypothesis by an ugly fact'. But in fact, while descriptively unrealistic, Krugman's model of target zones maintains its conceptual grip. A number of minor amendments to the theory have gone some way to reconciling it with the evidence while leaving its central ideas intact. The theory makes many clearly unrealistic assumptions. Two changes that have been particularly important are allowing for imperfect credibility of the target zone and allowing for intra-marginal intervention. Intra-marginal intervention in particular alters the process driving the fundamentals and can cause the theoretically predicted distribution of the exchange rate within the zone to have the empirical humped shape. Expectations that a zone is not fully credible and that the authorities might adjust the central parity rather than defend the zone also have the effect of reducing the curvature of the S-shaped curve and bringing it closer to the 45-degree line that would prevail under free floating.

With the empirical failure of the theory and the disappearance of the most prominent practical example when the euro replaced the EMS in 1999, interest in target zones has subsided. Nevertheless, John Williamson (1998) emphasizes the empirical observation that within a target zone the forward rate responds less than one-for-one with a change in the spot rate, whereas the same is not true of a floating exchange rate, as an indication that even an imperfectly credible target zone exerts a stabilizing influence on exchange rates. He has proposed looser arrangements for exchange rates, such as 'crawling bands' and 'monitoring bands', in which the band is defined around an equilibrium real exchange rate (his concept of the fundamental equilibrium exchange rate), which naturally implies a central nominal exchange rate that crawls over time. While a crawling band involves a commitment to keep the exchange rate within a wide announced band, a monitoring band involves a weaker commitment. A number of countries used crawling bands during the 1990s, including Chile, Colombia, Israel, Indonesia, Ecuador, and Russia. The IMF (2006) reports that by the end of 2005 no countries were using crawling bands. Several countries were using 'pegged exchange rates within horizontal bands', mostly countries in ERM II, the revised form of the Exchange Rate

Mechanism of the former EMS: Cyprus, Denmark, the Slovak Republic, Slovenia, and Hungary; with Tonga alone outside ERM II.

See Also

- ▶ Bretton Woods System
- ▶ Bubbles
- ▶ Currency Crises
- ▶ European Monetary Union
- ▶ Exchange Rate Volatility
- ▶ Irreversible Investment

Bibliography

Flood, R.P., A.K. Rose, and D.J. Mathieson. 1991. An empirical examination of exchange rate target zones. *Carnegie-Rochester Series on Public Policy* 35: 7–65.

Giavazzi, F., and A. Giovannini. 1989. *Limiting exchange rate flexibility: The European monetary system*. Cambridge, MA: MIT Press.

IMF (International Monetary Fund). 2006. De facto classification of exchange rate regimes and monetary policy frameworks. 31 December. Online. Available at http://www.imf.org/external/np/mfd/er/2005/eng/1205.htm. Accessed 23 Oct 2006.

Krugman, P.R. 1991. Target zones and exchange rate dynamics. *Quarterly Journal of Economics* 106: 669–682.

Svensson, L.E.O. 1992. An interpretation of recent research on exchange rate target zones. *Journal of Economic Perspectives* 6(4): 119–144.

Williamson, J. 1998. Crawling bands or monitoring bands: How to manage exchange rates in a world of capital mobility. *International Finance* 1: 59–79.

Exchange Rate Volatility

Roberto Rigobon

Abstract

Exchange rate volatility is at the forefront of academic, policy, and market participant discussions in developed and developing economies. This article reviews the benefits and cost of exchange rate variations, and its implications for the economy. It also summarizes the general understanding that exists regarding the relationship between real and nominal exchange rates.

Keywords

Aggregation bias; Bretton woods system; Consumer price index (CPI); Contagion; European monetary union; Exchange rate dynamics; Exchange rate mechanism (EU); Exchange rate regimes; Exchange rate volatility; Financial market contagion; Fixed exchange rates; Flexible exchange rates; Inflation targeting; Monetary base; Nominal exchange rates; Price revelation; Protection; Purchasing power parity; Real exchange rates; Speculation; Sticky prices; Uncovered interest parity

JEL Classifications

F31; F4

The volatility of the exchange rate is perhaps one of the most studied topics in international economics; from the real exchange rate to the nominal exchange rate, from the theories of exchange rate volatility to the predictability of exchange rate movements, from the explanations of the volatility to its implications, and from the independence of exchange rates around the world to contagion, it can be argued that all dimensions have been heavily scrutinized in the literature. (As it is virtually impossible to cite all the relevant sources, I will mostly refer to the classics and to relevant surveys in the literature.)

Why is so much attention paid to exchange rate volatility? For a start, exchange rate volatility has almost unanimously been seen as a *bad thing* by policymakers, practitioners, and academics – whatever 'bad' or 'thing' means. Indeed, an important policy objective in the recent past, for several emerging and developed nations, has been exchange rate stability. The road to this stability, however, has been long and with many bumps and pitfalls. For instance, one of the goals of the Bretton Woods system was to control the

excessive exchange rate variability that surfaced during the interwar period – it was unsuccessful, and in the early 1970s the world abandoned fixed exchange rates. After teasing the world with flexible rates for less than a decade, the European nations moved again to controlling their exchange rates with the Exchange Rate Mechanism (ERM) in early 1980s. It was a more comprehensive set of rules devoted to enhance cooperation and coordination among the members with the purpose of reducing exchange rate variability. It was a good idea, but unfortunately the system did not last, and it collapsed in the early 1990s. Europe stirred itself in 1999 to adopt an even stronger set of rules to foster even more cooperation and coordination – the European Monetary Union (EMU). Will it last? Too close to call, yet.

Emerging markets show even stronger variability of exchange rates, and more frequent failures of fixed regimes. Countries usually embark on periods of controlled nominal exchange rates in the hope that they can achieve stability, both internal and external. Those efforts rarely last, and most such countries are forced to devalue. In the end, the search for the Holy Grail of exchange rate stability looks more like an electrocardiogram than a smooth path.

Interestingly, this path has been frustrating enough that some countries have given up their exchange rate stabilization objective. The recent shift towards inflation targeting by many central banks is an indication that today they are assigning a smaller weight to exchange rate stabilization in the design of monetary policy than in the past. However, as the present global imbalances debate on the sustainability of the US current account deficit and the Asian and European current account surpluses shows, the exchange rate is still at the forefront of the concerns of policymakers, practitioners and academics.

Several questions come to mind. First, from the economics point of view, which volatility is the relevant one: nominal or real? Second, what are the costs and benefits of exchange rate stability? Third, what causes exchange rate instability, and how can it be controlled? The following sections answer these questions.

Nominal vs Real Exchange Rate Volatility

If we were to analyse the impact of higher exchange rate volatility on growth or trade, the first question would be whether the nominal or the real exchange rate should be considered. In practice, nominal and real exchange rates are closely intertwined. In his seminal paper, Mussa (1986) reports one of the most robust facts in international economics: the nominal and real exchange rates move almost one to one; periods of nominal exchange rate stability are always associated with periods of stable real exchange rates, while periods of nominal exchange rate instability are accompanied by excessive real exchange rate variability. Furthermore, a country that shifts from a fixed nominal exchange rate to a flexible nominal exchange rate experiences an increase in the variance of both nominal and real exchange rates. This is particularly true at shorter frequencies (quarterly or monthly). This evidence points to two important lessons. First, prices are sticky in the short run, so that nominal and real exchange rates are driven by the same factor, namely, shocks to the nominal exchange rate. Second, demand shocks – or nominal shocks – govern the short-run dynamics of the nominal exchange rate.

An active recent area of research has been the collection of empirical evidence regarding the degree of stickiness of prices using micro data. The seminal paper in this area is by Bils and Klenow (2004), who study price stickiness in the items used to construct the consumer price index (CPI) for the United States. They find that the median stickiness is four to five months. By contrast, Alvarez et al. (2005) find a degree of stickiness of about a year in European items constituting the CPI. Furthermore, for the US prices used to construct the import and export indexes, Gopinath and Rigobon (2006) find that the degree of stickiness is greater than a year. The significance of sticky prices and the degree of stickiness continue to be an important area of research. The recent evidence suggests a substantial degree of stickiness, meaning that short-run movements of the nominal rate will be transmitted one-to-one to the real exchange rate.

Finally, this short-run volatility pans out in the long run, at least in terms of real exchange rate deviations. As summarized by Rogoff (1996), the consensus view in the profession is that purchasing power parity holds in the long run. The average half-life of real exchange rate deviations from trend is around three to four years. More recent evidence has challenged this view. Imbs et al. (2005) have found that at the sectoral level the mean reversion is even stronger (half-lives of around a year). They argue that the very strong persistence of the aggregate real exchange rate measures is due to aggregation bias. The jury is still out on this matter. Nevertheless, it is clear that fluctuations of the real exchange rate are short-lived – where short is to be determined.

In summary, short-run fluctuations of the nominal and real exchange rates come in tandem, and therefore, from the policy point of view, stabilizing one is equivalent to stabilizing the other.

Implications of Exchange Rate Volatility

What are the advantages or disadvantages of exchange rate volatility? Because the variance is regime dependent – meaning fixed exchange rate regimes will have lower variance and flexible regimes will have higher variance – answering this question is closely linked to the advantages and disadvantages of the different exchange rate regimes. To simplify the discussion, let us concentrate on the two extremes: fixed and flexible.

As highlighted by Friedman (1953), one of the main advantages of flexible exchange rates is that they allow an independent monetary policy. Under fixed exchange rates, shifts in the demand for currency imply portfolio recompositions that are met instantaneously by changes in the international reserves. In this sense, when exchange rates are fixed an expansion of the monetary base implies an immediate loss of reserves. Therefore, if the fixed exchange rate regime is credible, the interest rate differential between the domestic rate and the international rate has to be zero, or small, limiting the scope of monetary policy management.

Under a flexible regime, the appreciation or depreciation of the exchange rate, and the risk premium implied by those movements, entails the possibility that domestic and international interest rates differ. Hence, in the presence of a supply shock and sticky prices, flexible nominal rates allow for an easier adjustment of the external account and unemployment. Conversely, in a fixed exchange rate regime, a negative shock implies that the economy would need large fluctuations in real activity to produce the desired price and wage changes – that is, under fixed exchange rates a negative shock implies a large increase in unemployment to achieve a drop in the real wage.

Finally, as has been argued by Tornell and Velasco (1995), flexible regimes react immediately to any shock – hence, irresponsible fiscal policy (for example) is felt immediately through a nominal depreciation, instead of a prolonged decline in international reserves, and an ultimate collapse.

In all these dimensions, a flexible exchange rate – and therefore a volatile exchange rate – seems to be superior to a fixed exchange rate. Can the exchange rate be excessively volatile, and generate costs to the economy?

Indeed, it is possible that exchange rates are excessively volatile. The first explanation was provided by Dornbusch, and the second was advanced by Mussa.

As Rudiger Dornbusch used to say in most seminars at Massachusetts Institute of Technology, 'exchange rates are determined in stock markets', emphasizing the asset price view of the exchange rate he shared, and developed. Indeed, one of the most influential theories in international economics was Dornbusch's (1976) *overshooting* theory of the exchange rate. (In my view, his paper is the most influential idea in international open macroeconomics since the mid-1970s.) The simple intuition of the overshooting model is that the nominal exchange rate has to satisfy two equations or restrictions: first, an asset equation (in this case the uncovered interest rate parity); and second, a 'real' equation (long-run purchasing power parity). In a world of fixed prices in the short run, changes in the stance of monetary policy create excessive volatility of the nominal exchange rate – where excess is here defined as

overshooting. For instance, after a loosening of monetary policy, we know that in the long run the exchange rate has to depreciate. However, the current increase in the money supply implies a reduction of the nominal interest rate. In order to satisfy the uncovered interest rate parity while prices adjust, the exchange rate should be appreciating (instead of depreciating). Therefore, the exchange rate has to depreciate excessively today, so that it can appreciate on the path towards the new equilibrium. This theory implies that exchange rates in the short run are more volatile than fundamentals.

The second theory explaining excessive exchange rate volatility is based on the fact that speculation in the market can be destabilizing. In standard models with flexible prices, speculation in the market is usually good because price movements will reveal the fundamentals that private agents are using to value the asset. However, if there is no full price revelation, speculation might destabilize the exchange rate market. This has been called in the literature the *magnification effect* (Mussa 1976). The reason is that expectations might not imply that speculators purchase assets when prices are going down, and sell them when prices are going up. Their actions depend on the properties behind the exchange rate stochastic process, and how agents form their expectations. Therefore, it is possible that excessive volatility of the nominal exchange rate is due to shocks to expectations, and not to fundamentals. This *noise* in the nominal exchange rate plays no role in the adjustment process.

It is possible to argue that excess exchange-rate volatility has real costs. The change in the real exchange rate has implications for production allocation – tradable versus non-tradable sectors, on hedging, on investment, and so on. It has been documented that a more volatile exchange rate reduces the amount of trade (Frankel and Wei 1993), increases the pressures toward protectionism (see Dornbusch and Frankel 1987, for a survey of the protectionist forces that appeared during the 1980s in the United States), increases the degree of persistence of inflation or deflation, so slowing the adjustment of the real exchange rate (Obstfeld 1995), and reduces the development of the financial sector (Aghion et al. 2005; Eichengreen et al. 2003). It is important to highlight that this empirical evidence is mostly suggestive. There are problems of simultaneous equations and omitted variables in answering any of these questions. Furthermore, there are no available instruments for the exchange rate volatility – leaving the literature mainly reporting correlations as opposed to causal relationships. Indeed, this is perhaps an area of research that will need to be revived in the future. However, the 'consensus' points to the costs emphasized before. Interestingly, even if the academic literature does not have a clear view on whether the volatility of the exchange rate is good or bad for the economy, it seems that policymakers and practitioners agree that exchange rate volatility is costly.

Dealing with Exchange Rate Volatility

Is exchange rate volatility a policy choice? Can exchange rate volatility be controlled? In other words, if we were to accept that exchange rate volatility is detrimental to the economy because it is excessive, then what is the policy advice? How can it be reduced?

These questions are related to the theories behind exchange rate determination. For example, if fundamentals determine the exchange rate, is should be the case that the exchange rate, and its volatility, can be controlled, or at least ameliorated, by changing the fundamentals. On the other hand, if exchange rates and their volatility are unrelated to fundamentals, then very little can be done from the policy point of view.

A substantial literature reports a 'disconnect' between exchange rates and fundamentals in the short run. The seminal paper by Messe and Rogoff (1983) argues that the best predictor of tomorrow's exchange rate is today's exchange rate. They show that in the short run the random-walk assumption beats most models of the exchange rate. Recently, however evidence has emerged that models of the exchange rate have performed better than the random walk for medium- and long- run horizons. (See Chin and Messe 1995, for the first paper in this area, and see the

subsequent papers by Menzie Chin on the subject.) This disconnect implies that very little can be done in the short run to control for the exchange rate volatility – other than fixing it. In fact, Evans and Lyons (2002) show that 70 per cent of the variation of the nominal exchange rate in the short run is explained by order flows in the market – meaning that market microstructure factors dominate the fluctuations of the exchange rate.

More interesting is the fact that in small open economies the exchange rate is governed by the fundamentals of other countries. The very well-known anomaly called 'contagion' implies that excess exchange rate volatility is affected by crises experienced by trading partners and in countries that share financial linkages. (See Forbes and Rigobon 2001, 2002, and Kaminsky et al. 2003, for detailed surveys of the empirical literature, and see Pavlova and Rigobon 2006, 2007, for the theories of contagion.) These are rarely subject to the influence of policymakers, and very little can be done in this respect.

In summary, the short-run volatility of a flexible exchange rate cannot be controlled by policymakers. The short-run fluctuations depend either on market participants' views or on other countries' fundamentals. In both cases, the only response open to the monetary authorities is to move toward a fixed regime – which is indeed what most countries do. This is a very imperfect policy measure, but unfortunately the only one available. Hence, there are two types of country. In the first, market participants' views are very volatile, or are subject to massive external shocks; these countries end up adopting fixed exchange rate regimes. For those countries the short-run fluctuation of the exchange rate is so costly that giving up monetary policy seems a minor issue. The second type of country consists of those that can bear the cost of the short-run exchange rate volatility, and have been moving towards a flexible rate – and towards inflation targeting. In fact, in the life of a single country it is easy to think of times in which nominal shocks dominate the economy (Argentina in the 1990s), which they responded to by fixing the exchange rate. However, after they introduced a fixed exchange rate the economy was mostly governed by supply shocks (high unemployment) and needed a flexible regime, which Argentina also implemented in 2002.

Final Remarks

Exchange rate volatility is one of the most important policy matters in developed economies, and possibly *the* topic in developing ones. The rhetoric of public policy is that exchange rate volatility is costly. In this article we have tried to understand the economic forces behind this claim.

First, we observed that in the short run the nominal exchange rate and the real exchange rate move in tandem. This is mainly due to the presence of sticky prices and real rigidities. This means that the discussion about volatility must embrace both the real and the nominal exchange rate.

Second, we have reviewed some of the evidence pointing out that exchange rate volatility is indeed costly for the economy. Investment is lower, growth is lower, real interest rates are higher, there are costly resource allocations between tradable and non-tradable sectors, and so on. This evidence, however, is not conclusive. There are econometric challenges, such as simultaneity and omitted variable biases, that have not yet been overcome. Nevertheless, it seems fair to say that the consensus points out to exchange rate volatility being costly to economies.

Third, we have observed that, even if we were to accept that exchange rate volatility is costly, there is very little that policymakers can do. There is a tremendous disconnect between exchange rates and fundamentals in the short run, and therefore policymakers are left with only one instrument to deal with exchange rate volatility: to fix it. That implies that only countries that face large demand shocks end up fixing their exchange rates, while most countries have been moving towards more flexible regimes.

See Also

► Purchasing Power Parity

Bibliography

Aghion, P., G. Angeletos, A. Banerjee, and K. Manova. 2005. *Volatility and growth: Credit constraints and productivity-enhancing investment*. Working Paper No. 11349. Cambridge, MA: NBER.

Alvarez, L.J., E. Dhyne, M. Hoeberichts, C. Kwapil, H. Le Bihan, P. Lunnemann, F. Martins, R. Sabbatini, H. Stahl, P. Vermeulen, and J. Vilumen. 2005. *Sticky price in the Euro area: A summary of new micro evidence*. Working Paper No. 563. Frankfurt am Main: European Central Bank.

Bils, M., and P. Klenow. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112: 947–985.

Chin, M., and R. Messe. 1995. Banking on currency forecasts: How predictable is change in money? *Journal of International Economics* 38: 161–178.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84: 1161–1176.

Dornbusch, R., and J. Frankel. 1987. Macroeconomics and protection. In *US trade policies in a changing world economy*, ed. R. Stern. Cambridge, MA: MIT Press.

Eichengreen, B., R. Hausmann, and U. Panizza. 2003. *Currency mismatches, debt intolerance and original sin: Why are not the same and why it matters*. Working Paper No. 10036. Cambridge, MA: NBER.

Evans, M., and R. Lyons. 2002. Order flow and exchange rate dynamics. *Journal of Political Economy* 110: 170–180.

Forbes, K., and R. Rigobon. 2001. Contagion in Latin America: Definitions, measurement, and policy implications. *Economia* 1(2): 1–46.

Forbes, K., and R. Rigobon. 2002. No contagion, only interdependence: Measuring stock market co-movements. *Journal of Finance* 57: 2223–2261.

Frankel, J., and S.-J. Wei. 1993. Trade blocs and currency blocs. In *The monetary future of europe*, ed. G. de la Dehesa. London: CEPR.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*. Chicago: University of Chicago Press.

Gopinath, G., and R. Rigobon. 2006. *Sticky borders*. Working Paper No. 12095. Cambridge, MA: NBER.

Imbs, J., H. Mumtaz, M. Ravn, and H. Rey. 2005. PPP strikes back: Aggregation and the real exchange rate. *Quarterly Journal of Economics* 120: 1–43.

Kaminsky, G.L., C.M. Reinhart, and C.A. Vegh. 2003. The unholy trinity of financial contagion. *Journal of Economic Perspectives* 17(4): 51–74.

Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–491.

Messe, R., and K. Rogoff. 1983. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics* 14: 3–24.

Mussa, M. 1976. The exchange rate, the balance of payments, and monetary and fiscal policy under a regime of controlled floating. *Scandinavian Journal of Economics* 78: 229–248.

Mussa, M. 1986. Nominal exchange rate regimes and the behavior of real exchange rates: Evidence and implications. In *Real business cycles, real exchange rates, and actual policies*, ed. K. Brunner and A. Meltzer. Amsterdam: North-Holland.

Obstfeld, M. 1995. International currency experience: New lessons and lessons relearned. *Brookings Papers on Economic Activity* 1995(1): 119–220.

Obstfeld, M., and K. Rogoff. 2005. Global current account imbalances and exchange rate adjustments. *Brookings Papers on Economic Activity* 2005(1): 67–123.

Pavlova, A., and R. Rigobon. 2006. *The role of portfolio constraints in the international propagation of shocks*. Mimeo: Massachusetts Institute of Technology.

Pavlova, A., and R. Rigobon. 2007. Asset prices and exchange rates. *Review of Financial Studies* 20: 1139–1181.

Rogoff, K. 1996. The purchasing power parity puzzle. *Journal of Economic Literature* 34: 647–668.

Tornell, A., and A. Velasco. 1995. Fiscal discipline and the choice of exchange rate regime. *European Economic Review* 39: 759–770.

Exchange Rates

Robert Z. Aliber

The large movements in the price of the US dollar in terms of the German mark, the Japanese yen, the British pound, and various other currencies since the breakdown of the Bretton Woods system in the early 1970s again raises the question of how exchange rates are determined. This question tends to be dormant when the major countries peg their currencies – and then reappears when their currencies float. Approaches to explaining the movement of the exchange rate must recognize that the range of variation in the price of the US dollar in terms of the currencies of various other countries has been substantially larger than the contemporary change in the differential between the increase in the US price level and the increase in the price levels in these other countries. Moreover, deviations between market exchange rates and real (or price-level adjusted) exchange rates have been substantially larger with

the floating exchange rate system than with the pegged exchange rate system of the 1950s and the 1960s. A second observation is that at times countries with strong and appreciating currencies have had large trade deficits (the United States in the early 1980s) and the countries with weak and depreciating currencies have had large trade surpluses (the United States in the late 1970s) – even though large trade deficits frequently are associated with weak or depreciating currencies, and large trade surpluses with strong or appreciating currencies. A third observation is that the range of variation in the price of the US dollar in terms of various foreign currencies does not appear to have declines during the first decade of experience with the floating exchange rate system. A fourth observation is that the differences in interest rates on comparable assets denominated in the US dollar and various foreign currencies have proven to be poor predictors of the rate of change of the price of the US dollar in terms of each of these currencies. Similarly, forward exchange rates have not proven to be effective predictors of future spot exchange rates at the maturity of the forward contracts.

The experience with floating exchange rates has proved very different from the experience with the Bretton Wood system of pegged exchange rates in the 1950s and the 1960s. Then countries devalued their currencies when their trade deficits were excessively large; changes in currency parities generally were consistent with the changes in international competitiveness. The countries which devalued their currencies, like France in 1959 and again in 1969 and Great Britain in 1967, had experienced higher rates of inflation than their major trading partners. And the countries which revalued their currencies, like Germany in 1961 and again in 1969, generally experienced lower rates of inflation than their major trading partners.

The next section discusses five approaches toward explaining the level of the exchange rate and changes in exchange rates. Then a disequilibrium toward analysing exchange rate movements is contrasted with an equilibrium approach. Then the relation between the determinants of exchange rates under a floating exchange rate system is compared with the determinants of the exchange rate under a pegged exchange rate system.

Approaches Toward Modelling the Determination of Exchange Rates

Most of the approaches toward explaining changes in the exchange rate were developed to explain phenomena at particular times. Five approaches are distinguished: purchasing power parity, elasticities, absorption, portfolio balance, and the asset market approach. The purchasing power parity approach is identified with Cassel, who sought to develop a way for European governments to determine the equilibrium values for their currencies if they again were to peg them to gold after World War I. During the war, inflation rates varied extensively among countries. Either the countries with the more rapid inflation would have to accept large declines in their price levels, or they would be obliged to peg their currencies to gold at new parities that would reflect that their inflation rates had been higher than those of their major trading partners. Cassel's insight was that changes in exchange rates should conform to differences in national inflation rates – in effect an extension of the arbitrage proposition known as the Law of One Price from individual goods to national market baskets of goods (traded) and services (non-traded). The terms *undervaluation* and *overvaluation* are the layman's expression that the value for the exchange rate seems inconsistent with the relationship between the domestic price level and the price levels in the major trading partners. Subsequent analysis has been directed at whether the equilibrium exchange rate should be inferred from absolute price levels or whether instead the exchange rate should be based on changes in price levels from data when the exchange market was in equilibrium should provide the basis for determining the new equilibrium exchange rate.

The elasticities approach to the determination of the exchange rate developed the 1930s in response to the observation – or at least to the stylized fact – that countries which devalued their currencies were not successful in increasing their exports relative to their imports. The devaluation improved competitiveness in that the price of domestic goods fell relative to the price of foreign goods, but the improvement in

competitiveness did not lead to the desired reduction in the trade deficit. The explanation was that the spending on imports might increase if domestic demand were price inelastic and export receipts might decline if foreign demand for domestic goods were price inelastic. This 'elasticity pessimism' view led to the conclusion that changes in exchange rates would prove ineffective in improving the trade balance, which was formalized in the Marshall–Lerner condition (that a devaluation would reduce the trade deficit if the sum of the elasticities is greater than one). An alternative interpretation for the observation that the devaluation of one country's currency would not reduce its trade deficit in the 1930s was that the subsequent devaluations of the currencies of its trading partners effectively neutralized its own devaluation.

The absorption approach to the determination of the exchange rate was developed in the period after World War II to highlight that changes in exchange rates would not lead to a permanent improvement in the trade balance unless the devaluating country adopted a sufficiently contractive monetary and fiscal policy. This approach followed the Keynesian tradition that the trade balance was the residual between domestic consumption and domestic production; a trade deficit occurred when domestic consumption exceeded domestic production, which meant that imports exceeded exports. Unless consumption declined relative to production as domestic currency was devalued, the trade deficit would persist. Thus a devaluation might be necessary to reduce a trade deficit; if excess demand remained after the devaluation, then domestic price level would rise, and the improvement in competitiveness effected by the devaluation would be negated by the subsequent increase in consumption and in imports.

Both the absorption approach and the elasticities approach provide explanations of why a devaluation would be ineffective in improving the trade balance in terms of the levels of demand. Thus the elasticities approach highlighted a deficiency of demand associated with the Great Depression, while the absorption approach reflects excess demand of the years immediately after World War II. In both cases, however, the equilibrium exchange rate was determined by the need to have national price levels aligned so that there would be equilibrium in the goods market; the change to price-level competitiveness was necessary for a permanent improvement in the trade balance, but not a sufficient condition.

The Bretton Woods decades were marked by infrequent changes in exchange parities of the industrial countries. Inflation rates in most countries were low. For most of this period, the United States incurred payments deficits; the problem was to explain the persistence of the US payments deficit despite a variety of measures adopted to reduce the imbalance. The payments surpluses of other industrial countries were explained by their demand for international reserves, or by their demand for money. The portfolio balance approach emphasized that payments balances and hence the exchange rate, reflected trade in securities as well as trade in goods. Trade in securities would involve a stock adjustment in the volume of foreign securities owned by domestic residents. The Monetary Theory of the Balance of Payments emphasized that payments surpluses and deficits reflected imbalances between the demand for money in each country and the supply of reserves that results from the monetization of domestic assets; if the demand for money increased more rapidly than the supply based on the expansion of the domestic assets owned by the central bank, then the country would realize a trade and payments surplus, since the supply of goods will exceed the demand. The inflow of gold and foreign exchange would lead to increases in the assets of the central bank, and thus lead to an increase in the money supply. Both the portfolio balance approach and the monetary theory placed payments surpluses and deficits in a general equilibrium framework. The monetary approach was generally mute on whether the payments surplus reflected the trade account or the capital account. In contrast, the portfolio balance approach could explain the US payments in terms of the desire of residents of other countries to borrow long and lend short in their transactions with the United States.

The observations about the wide variation in the price of national currencies since the early 1970s are similar to the observations about the

movements in exchange rates in the early 1920s. In both cases, the movements of exchange rates were much larger than the movement that would be inferred from the changes in the relationship among national price levels. In both periods, countries with relatively high inflation rates experienced a significantly more rapid reduction in the foreign exchange value of their currencies than would be inferred from the relative price level movements alone.

The dominant explanation for the large variations in market exchange rates derives from Irving Fisher's observation in *The Theory of Interest* that the interest rate differential between bonds payable in gold and bonds payable in rupees or silver reflected the anticipated rate of change in the price of gold in terms of silver. Thus the current spot exchange rate at any moment is the anticipated spot exchange rate for various future dates discounted to the present by the differential between interest rates on domestic securities and interest rates on similar securities denominated in the foreign currency. If the current spot exchange rate differed substantially from the anticipated spot exchange rate adjusted for the interest rate differential, investors would have a virtually riskless profit opportunity. The implication is that the spot exchange rate changes whenever the anticipated spot exchange rate changes, or whenever the differential between interest rates on comparable securities denominated in the domestic currency and foreign currency changes.

The Asset Market Approach to the Exchange Rate places the Fisherian observation in a general equilibrium framework. The exchange rate is the price of two national monies. Changes in the exchange rate reflect changes in the demand for securities denominated in each currency relative to the supply. The changes in the exchange rate that occur to obtain equilibrium in the asset market may induce disequilibrium in the goods market; the goods produced by the countries subject to capital outflows will become undervalued. In effect, the capital outflow can occur only if the country can generate a current account surplus. The asset market approach slights the role of trade in goods in determining the value of the exchange rate today, on the presumption that the daily volume of transactions in assets across national borders is so much larger than the volume of trade in goods. However, the anticipated exchange rate may reflect the value that would lead to goods market equilibrium at the anticipated price levels during future years.

The exchange rate necessary to achieve asset market equilibrium leads to disequilibrium in the goods market. The anticipated spot exchange rate reflects the value that will clear the goods markets at some future date; this anticipated value may be extrapolated from current or recent movements in the domestic and foreign inflation rate. In this way, changes in the inflation rate can have a major impact on the current spot exchange rate as the revised anticipated values are discounted to the present.

A Disequilibrium Approach to Exchange Rate Determination

Sudden large movements of exchange rate during particular episodes have been explained in terms of extrapolation by investors of the future exchange rate from the direction of movement in the exchange rate in the recent past. For a while, at least, some participants in the exchange market rely on a 'follow-the-leader' approach; changes in exchange rates thus reflect a bandwagon effect. Hence there may be 'speculative bubbles' in the exchange rate. Momentum models of exchange rate forecasting are based on this view. In such cases, the exchange rates may move away from an equilibrium value, and eventually these sharp movements will be reversed.

The approach toward explaining changes in exchange rates raises the question whether the foreign exchange market is efficient, or whether instead period-to-period movements in the exchange rate are serially correlated. The serially correlated movements in exchange rates could explain why the exchange rate tends to overshoot the ultimate equilibrium. In a few brief episodes, exchange rate movements appear serially correlated.

The momentum approach toward the determination of the exchange rate can be reconciled with

the Asset Market Approach in that new information about inflation rates may lead to sharp revisions in anticipated exchange rates and, to a lesser extent, in interest rates. Frequently this new information will develop over a period of weeks and months; each revision will lead to a new value for the current spot exchange rate. If the trend-like movement is sufficiently strong, then momentum-based approaches to forecasting exchange rates may be triggered, and the exchange rate movement may become abrupt, and overshoot its ultimate equilibrium.

Exchange Rate Determination – Pegged Rate Periods and Floating Exchange Rate Periods

The amplitude and suddenness of movements in exchange rates under the floating rate period highlights the conditions necessary for the successful operation of a pegged exchange rate regime – and for a floating exchange rate system in which movements in spot exchange rates are gradual, and conform with differences in inflation rates. A pegged exchange rate system can be maintained if the anticipated exchange rates are more or less identical with the current spot exchange rate – because the authorities are committed to maintaining their parities, and manage their policies accordingly. And for most periods, this commitment was credible. In contrast, during the floating rate periods, the authorities generally have had no such commitment about a particular value for the future spot exchange rate.

The sharp movements in the price of the US dollar in terms of various foreign currencies during the floating exchange rates reflects that various types of shocks affect the anticipated spot exchange rates, domestic interest rate and foreign interest rates. Most of these shocks reflected changes in monetary policy in the United States and in other countries.

The number and magnitude of monetary disturbances has been substantially larger during periods associated with floating exchange rates. In part, this reflects that many of the moves to floating exchange rates occur in an inflationary environment: the early 1920s and the 1970s were both periods of divergent movement in national price levels. During the 1950s and the early 1960s, inflation rates were low and similar among the industrial countries; the anticipated price of foreign exchange more or less approximates the parity. Interest rates differed modestly among countries. Capital flows occur primarily in response to national differences in savings and investment rates. The exchange rate at the level necessary to maintain the approximate balance in trade accounts flow of capital were modest in size relative to flows of goods.

See Also

- Crawling Peg
- Fixed Exchange Rates
- Flexible Exchange Rates
- International Finance
- International Monetary Institutions
- Monetary Approach to the Balance of Payments

Bibliography

Alexander, S.S. 1952. Effects of a devaluation on a trade balance. *IMF Staff Papers* 2: 263–278.

Aliber, R.Z. 1973. The interest rate parity theorem: A reinterpretation. *Journal of Political Economy* 81(6): 1451–1459.

Aliber, R.Z. 1980. Floating exchange rates: The twenties and the seventies. In *Flexible exchange rates and the balance of payments*, ed. J.S. Chipman and C.P. Kindleberger. Amsterdam/Oxford: North-Holland.

Balassa, B. 1964. The purchasing-power-parity doctrine: A reappraisal. *Journal of Political Economy* 72: 584–596.

Bernstein, E.M. 1956. Strategic factors in balance of payments adjustment. *IMF Staff Papers* 5: 151–169.

Bilson, J.F. 1978. Rational expectations and the exchange rate. In *The economics of exchange rates*, ed. J.A. Frenkel and H.G. Johnson. Reading: Addison-Wesley.

Branson, W.H., and D.W. Henderson. 1985. The specification and influence of asset markets. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen. Amsterdam/Oxford: North-Holland.

Cassel, G. 1922. *Money and foreign exchange after 1914*. New York: Macmillan.

Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84(6): 1161–1176.
Fisher, I. 1930. *The theory of interest*. New York: The Macmillan Company.
Frenkel, J.A. 1976. A monetary approach to the exchange rate: Doctrinal aspects and empirical evidence. *Scandinavian Journal of Economics* 78(2): 200–224.
Frenkel, J.A. 1981. Flexible exchange rates, prices and the role of 'news': Lessons from the 1970s. *Journal of Political Economy* 89(4): 665–705.
Frenkel, J.A., and M. Mussa. 1985. Asset markets, exchange rates, and the balance of payments. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen. Amsterdam/Oxford: North-Holland.
Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.
Haberler, G. 1949. The market for foreign exchange and stability of the balance of payments: A theoretical analysis. *Kyklos* 3: 193–218.
Isard, P. 1977. How far can we push the 'law of one price'? *American Economic Review* 67(5): 942–948.
Mundell, R.A. 1960. The monetary dynamics of international adjustment under fixed and flexible exchange rates. *Quarterly Journal of Economics* 74: 227–257.
Mussa, M. 1979. Empirical regularities in the behavior of exchange rates and theories of the foreign exchange market. In *Policies for employment, prices, and exchange rates*, ed. K. Brunner and A.H. Meltzer. Amsterdam/Oxford: North-Holland.
Mussa, M. 1982. A model of exchange rate dynamics. *Journal of Political Economy* 90(1): 74–104.
Officer, L.H. 1976. The purchasing-power-parity theory of exchange rates: A review article. *IMF Staff Papers* 23(1): 1–60.

Exchangeability

David Draper

Abstract

Exchangeability is an invariance property of probability distributions that is central to the process of specifying Bayesian statistical models. Exchangeability judgements play a role in Bayesian modelling analogous to judgments in frequentist modelling that observable quantities may be regarded as realizations of independent identically distributed (IID) random variables. Judgements of conditional exchangeability (given the values of relevant covariates), when combined with Bayesian nonparametric modelling, provide a principled and rather general approach to Bayesian model specification that can lead to well-calibrated inferences and predictions; other approaches to achieving this goal include cross-validation and Bayesian model averaging.

Keywords

Bayes' th; Bayesian nonparametric methods; Bayesian statistics; Bernoulli distributions; Cumulative distribution functions; Dirichlet process priors; Exchangeability; Frequentist statistics; Markov chain Monte Carlo methods; Model averaging; Model specification; Model uncertainty; Pólya trees; Prior distributions; Random variables; Uncertainty

JEL Classification

C11

Definition A sequence $y = (y_1, \ldots, y_n)$ y of random variables (for $n \geq 1$) is (*finitely*) *exchangeable* if the joint probability distribution $p(y_1, \ldots, y_n)$ of the elements of y is invariant under permutation of the indices $(1,\ldots, n)$, and a countably infinite sequence $(y_1, y_2, \ldots)$ is (*infinitely*) *exchangeable* if every finite subsequence is finitely exchangeable.

The idea of exchangeability seems (Good 1965; Bernardo and Smith 1994) to be traceable back to Johnson (1924), who used the term *permutable*, and independently to Haag (1924). Other writers who made early use of the concept include Khintchine (1932); Fréchet (1943); Savage (1954), who called an exchangeable sequence *symmetric*; and Hewitt and Savage (1955). But the deepest implications of the idea are due to de Finetti (1930), who in (1938) still referred to exchangeable sequences as *equivalent*; by (1970) the word had been translated from the Italian as *exchangeable*, and since then usage has stabilized around this terminology under the influence of de Finetti and his translators.

The concept is important because it plays a fundamental role in the specification of statistical models from a Bayesian point of view. Following the example of Good (1950) by referring to You as a generic rational person making uncertainty assessments, suppose that You will in the future get to see a finite sequence $y = (y_1, \ldots, y_n)$ of binary observables; to illustrate the interplay between context and model, consider as an example the mortality outcomes (within 30 days of admission, say: 1 = died, 0 = lived) for a sequence of n patients with the same admission diagnosis (heart attack, say) at one particular hospital H, starting on the first day of next month. You acknowledge Your uncertainty about which elements in the sequence will be 0 s and which 1 s; suppose further that You find it natural (as in the Bayesian approach to statistics) to use random variables to quantify Your uncertainty. As de Finetti (1970) noted, in this situation Your fundamental imperative is to construct a *predictive* distribution $p(y_1, \ldots, y_n)$ that expresses Your uncertainty about the future observables, rather than – as is perhaps more common – to reach immediately for a standard family of *parametric* models for the y_i (that is, to posit the existence of a vector $\theta = (\theta_1, \ldots, \theta_k)$ of parameters and to model the observables by appeal to a family $p(y_i|\theta)$ of probability distributions indexed by θ).

Even though the y_i are binary, with all but the smallest values of n it still seems a formidable task to *elicit* from Yourself an n-dimensional predictive distribution p $(y_1, \ldots, y_n)$; it was while facing this challenge that de Finetti developed his version of the idea of exchangeability and its implications. As de Finetti observed, in the absence of any further information about the patients, You notice that Your uncertainty about them is exchangeable: if someone (without telling You) were to rearrange the order in which their mortality outcomes become known to You, Your predictive distribution would not change. This still seems to leave p-$(y_1, \ldots, y_n)$ substantially unspecified, but de Finetti (1930) proved a remarkable theorem which shows (in effect) that all exchangeable predictive distributions for a vector of binary observables are representable as mixtures of Bernoulli sampling distributions. More formally,

Theorem 1 (Representation of Exchangeable Predictive Distributions for Binary Observables [de Finetti 1930]) Suppose that You're willing to regard $(y_1, \ldots, y_n)$ as the first n terms in an infinitely exchangeable binary sequence $(y_1, y_2, \ldots)$; then, with $\overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$,

- $\theta = \lim_{n\to\infty}\overline{y}_n$ must exist, and the marginal distribution (given θ) for each of the y_i must be $p(y_i|\theta) = \text{Bernoulli}(\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$;
- $H(t) = \lim_{n\to\infty} P(\overline{y}_n \le t)$, the limiting cumulative distribution function (CDF) of the $\overline{y}_n$ values, must also exist for all t and must be a valid CDF, where P is Your joint probability distribution on $(y_1, y_2, \ldots)$; and
- Your predictive distribution for the first n observations can be expressed as

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} dH(\theta). \quad (1)$$

When (as will essentially always be the case in realistic applications) Your joint distribution P is sufficiently regular that H possesses a density (with respect to Lebesgue measure), $dH(\theta) = p(\theta)$, (1) can be written in a more accessible way as

$$p(y_1, \ldots, y_n) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n} p(\theta)\, d\theta, \quad (2)$$

where $s_n = \sum_{1=1}^{n} y_i = n\overline{y}_n$.

The interpretation of (2) provides a link with non-Bayesian statistical modelling, as follows. In the frequentist (repeated-sampling) approach to statistics, to bring probability into the picture it's necessary to tell a story in which the observable y_i are either literally a random sample from some population $\mathscr{P}$ or like what You would get if You took a random sample from $\mathscr{P}$. This is a somewhat awkward story to tell in the medical example above, because the patients whose mortality outcomes are $(y_1, \ldots, y_n)$ are not a random sample of anything; they're simply the exhaustive list of all patients arriving at hospital H (itself not randomly chosen), with heart attack as their admission

diagnosis, in a particular (not randomly chosen) window of time. In spite of this difficulty, the standard frequentist model (with the same information base as that assumed above) would define θ as the mortality rate in $\mathscr{P}$ (whatever $\mathscr{P}$ might be) and would treat the y_i as measurements on a random sample from $\mathscr{P}$ by regarding random variables Y_i (whose observed values are y_i) as independent and identically distributed (IID) draws from the Bernoulli (θ) distribution $p(Y = y_i) = \theta^{y_i}(1-\theta)^{1-y_1}$, which leads to the joint sampling distribution $p(Y_1 = y_1 \ldots, Y_n = y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{s_n}(1-\theta)^{n-s_n}$.Thus, in interpreting Theorem 1 (with reference to equation (2) above), in Your predictive modelling of the binary y_i You may as well proceed as *if*

- there is a quantity called θ, interpretable both as the marginal death probability $p(y_i = 1|\theta)$ for each patient and as the long-run mortality rate in the infinite sequence $(y_1, y_2, \ldots)$;
- conditional on θ, the y_i are IID Bernoulli (θ); and
- θ can be viewed as a realization of a random variable (this is of course how all unknown quantities are treated in the Bayesian paradigm) with density $p(\theta)$.

In other words, exchangeability of Your uncertainty about a binary process is functionally equivalent to assuming the simple Bayesian *hierarchical* model (see, for example, Draper 2007)

$$\begin{array}{ccc} \theta & \sim & p(\theta) \\ & \text{IID} & \\ (y_i|\theta) & \sim & \text{Bernoulli}(\theta) \end{array} \tag{3}$$

A number of points are worth noting.

First, exchangeability is not a property of the world; it's a judgment by You concerning Your uncertainty about the world. Two reasonable people, with different knowledge bases or different views on how that knowledge should be brought to bear on the issue at hand, may consider the same set of observables, and one may judge her uncertainty about those observables exchangeable while the other may not make the same judgement about his uncertainty (for example, if I know the gender of the patients in the medical example and You do not, and if there's evidence that the mortality rate for male and female heart attack patients differs by an amount that's large in clinical terms, then You may well judge Your uncertainty about the mortality outcomes exchangeable but I would be ill-advised to adopt an exchangeable model; see partial/conditional exchangeability below). This distinction between {the world} and {Your uncertainty about the world} is sometimes blurred by terminology – I might casually say 'These patients are exchangeable' when what I mean is 'my uncertainty about these patients is exchangeable, as far as mortality is concerned' – but failing to observe the distinction can lead to what Jaynes (2003) terms the *mind projection fallacy*, with undesirable consequences for clarity of thought.

Second, in both the frequentist and Bayesian modelling approaches it's helpful to employ a fiction involving random variables, but for different purposes: in the standard frequentist approach, You regard the yi as realizations of random variables (as a way to build a useful probability model), even though (in observational settings like the medical example above) no random sampling was performed to arrive at the observables; and in the Bayesian approach, You regard θ as random (as a way to make good predictions of observables), even though in both the frequentist and Bayesian approaches θ has the same logical status, as a fixed unknown constant.

Third, since, for any random variables X and Y for which the following symbols have meaning, the density of Y can be expressed as $p(y) = \int p(y|x)p(x)dx$ – in other words, Y can be modelled either directly or as a mixture of the conditional distribution $p(y|x)$ with $p(x)$ serving as a mixing distribution – the predictive distribution in Eq. (2) can be regarded as a mixture of Bernoulli sampling distributions with $p(\theta)$ as the mixing weights.

Fourth, mathematically $p(\theta)$ is just a mixing distribution, but (of course) statistically it has a more useful inpt. The second line of Eq. (3) defines the *likelihood function* $l(\theta|y) = c\,p(y|\theta)$ (an arbitrary positive constant c times the joint sampling distribution of the data vector y, reinterpreted as a function of θ for fixed y); this is where all the information about θ internal to the data-set y is stored, and – under the logic of *Bayes's theorem* – the first line of Eq. (3) defines $p(\theta)$ as the place where all the information about θ *external* to the data-set y is stored. It has become traditional to call this $p(\theta)$ a *prior distribution* for θ; this terminology is unfortunate (it sounds as though only information gathered before the data-set y arrives can go into $p(\theta)$, and this is not true), but it has been used for so long that it's unlikely it can be changed now. Equation (3) implies that (*a*) learning about θ on the basis of y can occur via Bayes's theorem: the *posterior distribution* $p(\theta|y)$, which combines the information about θ contained in the prior and the likelihood, is just a renormalized version of their product: $p(\theta|y) = c\,p(\theta)\,l(\theta|y)$, with c chosen so that $p(\theta|y)$ integrates to (1), and (*b*) predictive distributions for future data given past data may also readily be calculated (for example, for $1 < m < n$ the predictive for $y_{m+1}, \ldots, y_n$ based on $(y_1, \ldots, y_m)$ is

$$p(y_{m+1}, \ldots, y_n|y_1, \ldots, y_m) = \int_0^1 \theta^{s_m}(1-\theta)^{n-m(s_n-s_m)}p(\theta|y_1, \ldots, y_m)\,d\theta, \tag{4}$$

in which $s_m = \sum_{i=1}^{m} y_i$ and $p(\theta|y_1, \ldots, y_m)$ is the posterior distribution for θ based only on the first m observations.

Fifth, exchangeability evidently plays a role in Bayesian modelling that's somewhat analogous to the role of IID sampling in the frequentist approach, but exchangeability and IID are not the same: IID random variables are exchangeable, and exchangeable random variables are identically distributed, but they're not independent (for example, if You're about to observe a binary process whose tendency to yield a 1 is not known to You, and You judge Your uncertainty about future outcomes to be exchangeable, the information in the first n outcomes would definitely help You to predict outcome $(n+1)$; it's only when (somehow) the knowledge of the 'underlying' θ becomes available to You that there's no information in any of the outcomes to help predict any other outcomes – this situation might be summarized by saying that the past and the future become conditionally independent given the truth). Exchangeable observables are thus not IID, but they may often be usefully regarded as *conditionally* IID given a parameter vector θ, as in Eq. (3) above.

Sixth, some awkwardness arose above in the frequentist approach to modelling the medical data, because it was not clear what population $\mathscr{P}$ the data could be regarded as *like* a random sample from. This awkwardness also arises in Bayesian modelling: even though in practice You are only going to observe $(y_1, \ldots, y_n)$, de Finetti's representation theorem requires You to extend Your judgement of finite exchangeability to the countably-infinite collective $(y_1, y_2 \ldots)$, and this is precisely *like viewing* $(y_1, \ldots, y_n)$ as a random sample from $\mathscr{P}$. (Finite versions of de Finetti's representation theorem are available – for example, Diaconis and Freedman 1980 – which informally say that, if You're willing to extend Your judgement of exchangeability from $(y_1, \ldots, y_n)$ to $(y_1, \ldots, y_N)$ for $N > n$, the larger N is the harder it becomes for Your predictive distribution p(y1,..., yn) to differ by a large amount from something representable by Eq. (2) without violating the basic rules of probability.) The key point is that the difficulty arising from lack of clarity about the scope of valid generalizability from a given set of observational data is a fundamental scientific problem that emerges whenever purely observational data are viewed through an inferential or predictive lens, whether the statistical methods You use are frequentist or Bayesian.

E

The entire discussion so far has been in the context of binary outcomes y with no covariates; in practice, predictor variables x are also generally available, and extensions to non-binary data are evidently needed as well. An example involving a covariate arose in the discussion of the mind projection fallacy above: if, in the medical setting considered here, the gender x of the patients is available, and if this has a clinically meaningful bearing on mortality from heart attack, then it would be more scientifically appropriate to assert exchangeability separately and in parallel within the two gender groups {male, female}. With this in mind de Finetti (1938) defined the concept of *partial exchangeability*, which is also known as *conditional exchangeability* (Lindley and Novick 1981; Draper et al. 1993); with this newer terminology You would say that Your uncertainty about the mortality observables for these patients is conditionally exchangeable given gender. Conditional exchangeability is related to the notion, introduced by Fisher (1956), of *recognizable subpopulations*.

Suppose now that the observable You will measure on the patients in the medical example is a severity of illness score y_i, scaled as a continuous quantity from $-\infty$ to ∞, and return temporarily to the situation with no covariate information. As before Your uncertainty about the future yi values is (unconditionally) exchangeable, but now a representation theorem is needed for continuous real-valued outcomes; de Finetti (1937) supplied this as well.

Theorem 2 (Representation of Exchangeable Predictive Distributions for Continuous Observables [de Finetti 1937]) If You're willing to regard $(y_1, \ldots, y_n)$ as the first n terms in an infinitely exchangeable sequence $(y_1, y_2 \ldots)$ of continuous values on $\mathbb{R}$, then

- $F(t) = \lim_{n\to\infty} F_n(t)$ must exist for all t and must be a valid CDF, where F_n is the empirical *CDF* based on $(y_1, \ldots, y_n)$ (i.e., $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le t)$, in which $I(A)$ is the indicator function (1 if A is true, otherwise 0)), and the marginal distribution (given F) for each of the yi must be $(y_i|F) \sim \cong$;
- $G(F) = \lim_{n\to\infty} P(F_n)$ must also exist, where P is Your joint probability distribution on $(y_1, y_2, \ldots)$; and
- Your predictive distribution for the first n observations can be expressed as

$$p(y_1, \ldots, y_n) = \int_{\mathscr{F}} \prod_{i=1}^{n} F(y_i)\, dG(F), \qquad (5)$$

where $\mathscr{F}$ is the space of all possible CDFs on $\mathbb{R}$. Equation (5) says informally that exchangeability of Your uncertainty about an observable processing unfolding on the real line is functionally equivalent to assuming the Bayesian hierarchical model

$$\begin{array}{ccc} F & \underset{\text{IID}}{\sim} & p(F) \\ (y_i|F) & \sim & F, \end{array} \qquad (6)$$

where $p(F)$ is a prior distribution on $\mathscr{F}$.

With binary observables, Theorem 1 (which is evidently a special case of Theorem 2) focuses attention on θ, the underlying rate of 1 s in the population $\mathscr{P} = (y_1, y_2, \ldots)$ from which You're in effect regarding $(y_1, \ldots, y_n)$ as like a random sample; in the continuous case the analogous theorem focuses attention on F, the underlying CDF defined by $\mathscr{P}$. This makes Theorem 2 harder to implement in practice, because it's one thing to specify a prior distribution on a quantity $\theta \in (0, 1)$ and quite another to put a scientifically relevant prior distribution on the space $\mathscr{F}$ of all possible CDFs on the real line. Placing probability distributions on functions is the topic addressed by the field of *Bayesian nonparametric methods* (see, for example, Dey et al. 1998), an area of statistics that has recently moved completely into the realm of day-to-day implementation and relevance through advances (since the early 1990s) in Markov chain *Monte Carlo* (MCMC) simulationbased methods of computation (see, for example, Gilks et al. 1995). Two rich families of prior distributions on CDFs about which a wealth of practical experience has recently accumulated include (mixtures of) *Dirichlet process priors* (see, for example,

Ferguson 1973) and *Pólya trees* (see, for example, Lavine 1992).

As an example of the use of de Finetti's representation theorem for continuous outcomes, consider a randomized controlled trial or observational study with a treatment (T) and a control (C) group in which the outcome of interest y is modelled continuously on $\mathbb{R}$. A judgement of unconditional exchangeability of Your uncertainty about the y values for all subjects in the study would be equivalent to assuming that the T and C conditions had the same effect on the subjects, which (since the point of the study is presumably to see if this is true) would not be a good starting point; instead, in the absence of any other covariate information, it would be reasonable for You to model Your uncertainty about the y values as conditionally exchangeable given the indicator variable x that identifies which group each subject is in. With F_C and F_T as the underlying control and treatment CDFs and y_i^T as the observable for subject i in the treatment group (and similarly for y_j^C), a straightforward extension of Theorem 2 then leads to the following Bayesian nonparametric model for the observables (for $i = 1, \ldots n_T$ and $j = 1, \ldots, n_C$):

$$\left(y_i^T | F_C, F_T\right) \underset{\sim}{\text{IID}} (F_C, F_T) \quad \text{and} \quad \left(y_j^C | F_C, F_T\right) \underset{\sim}{\text{IID}} F_C. \tag{7}$$

A nonparametric joint prior can then be placed on (F_C, F_T) using either of the Dirichlet process prior or Pólya tree methodologies mentioned above, and an appropriate functional of (F_C, F_T) (such as the difference or ratio of the underlying treatment and control means) can be monitored in the MCMC simulation. Note that this model arose solely from exchangeability considerations and (a simple extension of) Theorem 2.

Model specification has been a vexing topic in both frequentist and Bayesian statistics throughout much of the last century. Referring both to the conditional exchangeability judgements and to choices made in specifying the prior on F_C and F_T in the example above as *structural assumptions*, a popular approach to model specification (practised with equal vigour by both frequentists and Bayesians since the work of Tukey 1962, and others on exploratory data analysis) involves (*a*) enlisting the aid of the data to conduct a search among possible structural assumptions, (*b*) choosing a single favourite structural specification S^*, and (*c*) pretending You knew all along that S^* was 'correct', even though it was arrived at via a data-driven search. From a Bayesian perspective this approach is clearly unsound, since it amounts to using the data to specify the prior distribution on the space $\mathscr{S}$ of all possible structural assumptions and then using the same data again to update the prior on $\mathscr{S}$; the result will often be inferences and predictions that are not well *calibrated*, with interval estimates that are not as wide as they need to be to fully acknowledge *model uncertainty*. *Bayesian model averaging* (Leamer 1978; Draper 1995), in which predictive distributions $p(y_f|y)$ for future observables y_f given past data y are computed by averaging over the model uncertainty uncovered by the search through $\mathscr{S}$ (rather than ignoring it), through calculations of the form

$$p\left(y_f|y\right) = \int_S p\left(y_f|y, S\right) p(S|y)\, dS,$$

can provide one principled, satisfying and rather general method for solving the problem of Bayesian model specification in a well-calibrated manner; methods based on *cross-validation* (see, for example, Stone 1974), in which (in effect) part of the data is used to specify the prior on $\mathscr{S}$ and the rest of the data is employed to update that prior, can provide another; and the approach illustrated above in the study with treatment and control groups – which combines conditional exchangeability judgments (driven by the context of the problem) with Bayesian nonparametric methods – can provide yet another.

See Also

- ▶ Calibration
- ▶ Model Averaging
- ▶ Model Uncertainty

Bibliography

Bernardo, J.M., and A.F.M. Smith. 1994. *Bayesian theory*. New York: Wiley.

de Finetti, B. 1930. Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Accademia dei Lincei* 4: 86–133.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut H. Poincaré* 7, 1–68. Reprinted in translation as: Foresight: its logical laws, its subjective sources. In Studies in Subjective Probability, ed. H.E. Kyburg and H.E. Smokler. New York: Dover, 1980.

de Finetti, B. 1938. Sur la condition d'équivalence partielle. *Actualités Scientifiques et Industrielles* 739. Reprinted in translation as: On the condition of partial exchangeability. In *Studies in Inductive Logic and Probability*, ed. R.C. Jeffrey. Berkeley: University of California Press, 1980.

de Finetti, B. 1970. *Teoria delle Probabilità*, 2 vols. Turin: Eunaudi. Reprinted in translation as de Finetti B. 1974–75. *Theory of Probability*, 2 vols. Chichester: Wiley, 1974–75.

Dey, D.D., P. Müller, and D. Sinha (eds.). 1998. *Practical nonparametric and semiparametric Bayesian statistics*. New York: Springer.

Diaconis, P., and D. Freedman. 1980. Finite exchangeable sequences. *Ann. Probab.* 8: 745–764.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society: Series B* 57: 45–97.

Draper, D. 2007. Bayesian multilevel analysis and MCMC. In *Handbook of multilevel analysis*, ed. J. de Leeuw and E. Meijer. New York: Springer.

Draper, D., J. Hodges, C. Mallows, and D. Pregibon. 1993. Exchangeability and data analysis (with discussion). *Journal of Royal Statistics Society, Series A* 156: 9–37.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1: 209–230.

Fisher, R.A. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Fréchet, M. 1943. *Les Probabilités Associées à un Système d'Événements Compatibles et Dépendents*, vol. 2. Paris: Hermann et Cie.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter (eds.). 1995. *Markov Chain Monte Carlo in practice*. New York: Chapman & Hall/CRC.

Good, I.J. 1950. *Probability and the weighing of evidence*. London: Charles Griffin.

Good, I.J. 1965. *The estimation of probabilities: An essay on modern Bayesian methods*. Cambridge, MA: MIT Press.

Haag, J. 1924. Sur un problème général de probabilités et ses diverses applications. *Proceedings of the International Congress of Mathematics (Toronto)* 2: 659–674.

Hewitt, E., and L.J. Savage. 1955. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* 80: 470–501.

Jaynes, E.T. 2003. *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Johnson, W.E. 1924. *Logic. III. The logical foundations of science*. Cambridge, UK: Cambridge University Press.

Khintchine, A.I. 1932. Sur les classes d'événements équivalents. *Mathematics of the USSR – Sbornik* 39: 40–43.

Lavine, M. 1992. Some aspects of Pólya tree distributions for statistical modeling. *The Annals of Statistics* 20: 1222–1235.

Leamer, E.E. 1978. *Specification searches: Ad Hoc inference with nonexperimental data*. New York: Wiley.

Lindley, D.V., and M.R. Novick. 1981. The role of exchangeability in inference. *The Annals of Statistics* 9: 45–58.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Stone, M. 1974. Cross-validation choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society: Series B* 36: 111–147.

Tukey, J.W. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33: 1–67.

Excise Taxes

James R. Hines Jr.

Abstract

Excise taxes are selective taxes on the sale or use of specific goods and services, such as alcohol and petrol. Over time, governments have relied less on excise taxes, though, as of 2007, excise taxes still contribute 12 per cent of total government revenues in OECD countries. In addition to generating needed revenue, excise taxes can control externalities and impose tax burdens on those who benefit from government spending. Rather more controversially, they also can be used to discourage consumption of potentially harmful substances (such as tobacco and alcohol) that individuals might over-consume in the absence of taxation.

Keywords

Excise tax incidence; Excise taxes; Externalities; Leisure; Optimal taxation; Pigouvian taxes; Ramsey model

JEL Classifications

H2

Excise taxes are selective taxes on the sale or use of specific goods and services, such as alcohol and petrol.

Excise taxes have existed for centuries and are widely used by governments today, at the beginning of the 21st century. The 20th-century spread of income taxation and value-added taxation reduced the significance of excise taxation as a source of government revenue, but most governments still collect sizeable taxes on petroleum products, tobacco products and alcohol. For example, in 2004 the US federal government collected 72 billion dollars in excise taxes, representing four per cent of its total tax revenues, of which petroleum taxes accounted for 33 billion dollars, or 45 per cent of total excises. The United States relies the least on excise taxes of the 30 wealthy nations that are members of the Organisation for Economic Co-operation and Development (OECD), whose excise tax collections in 2000 averaged 12 per cent of total government revenues (Hines 2007). As recently as 1969–71, excise taxes contributed 23 per cent of total tax collections of high-income countries, and 27 per cent of tax collections of developing countries (Cnossen 1977). Among OECD countries in 2000, those with the lowest incomes, the most centralized governments and the greatest openness to international trade had the highest ratios of excise tax collections to total government revenues (Hines 2007).

Excise taxes take the form either of specific taxes or of ad valorem taxes. A specific tax (or unit tax) is defined per unit of the taxed good or service, whereas an ad valorem tax is defined per sales value. Thus, the current US federal taxes of 0.184 dollar per gallon of petrol and 0.39 dollar per pack of cigarettes are specific taxes, while the US tax of 11 per cent of the value of bows and arrows and three per cent of the value of fish-finding sonar devices are ad valorem taxes. In competitive markets specific and ad valorem taxes have identical consequences, other than any differences stemming from compliance and enforcement. In imperfectly competitive markets the difference is more consequential, since ad valorem taxes automatically impose higher per-unit tax rates as firms restrict quantities to drive up prices. Hence, in imperfectly competitive markets, ad valorem taxes produce lower consumer prices and lower deadweight loss than do specific taxes raising the same revenue (Suits and Musgrave 1953; Delipalla and Keen 1992).

Four motivations underlie the use of most excise taxes. The first is revenue generation: excise taxes can produce significant government revenues, and may do so at lower political or economic cost than alternatives such as income taxation. The second motivation is the application of the benefit principle of taxation: excise taxes can be tailored to impose tax burdens on those who benefit from government services financed by excise taxes. Petrol fuel taxes are often justified as user fees for government-provided roads, and the tax on sonar devices is justified by government expenditures to maintain lakes and fisheries. The third motivation is control of externalities, which is the goal of a number of excise taxes on polluting substances, such as taxes on ozone-depleting chemicals. And the fourth motivation is that excise taxes may discourage consumption of potentially harmful substances (such as alcohol and tobacco) that individuals might over-consume in the absence of taxation.

Excise Tax Incidence

It is customary to think of the burden of excise taxes as being borne by consumers of taxed goods and services in the form of higher after-tax prices, but there is considerable scope for the shifting of tax burdens. In a simple competitive partial equilibrium setting, the burden of an excise tax depends on elasticities of demand and supply: if demand for a taxed good or service is elastic, and supply relatively inelastic, then the burden of an excise tax is borne by sellers, whereas buyers bear

the burden of a tax on a good or service with inelastic demand and elastic supply. Similar demand and supply elasticities imply equal sharing of excise tax burdens between buyers and sellers.

There are plausible circumstances in which consumers can bear more than 100 per cent of the burden of an excise tax, or alternatively, might actually benefit from the introduction of a tax. If the market for a good or service is imperfectly competitive, then, depending on the nature of competition and cost conditions, consumers may well bear more than 100 per cent of the burden of an excise tax (Delipalla and Keen 1992). Even with perfect competition, the nature of demand and supply spillovers among multiple markets in general equilibrium can produce anomalous outcomes, such as consumer prices that rise by more than the amount of excise taxes, or (after-tax) prices that even fall with the introduction of excise taxes (Hotelling 1932).

There is mixed evidence of excise tax incidence in practice. Poterba (1996) offers evidence that consumers bear the full burden of US excise and sales taxes in the form of higher prices, but Besley and Rosen (1999) find that many consumer prices rise by more than 100 per cent of excise and sales taxes imposed by US states and localities.

One concern frequently expressed about excise taxation is the potential regressivity of the resulting tax burdens. Excise tax rates do not rise with consumption in the way that income tax rates rise with income; furthermore, since the poor spend higher fractions of their incomes than do the wealthy, taxes based on expenditure rather than income may put greater relative burdens on low-income individuals. This second consideration is considerably diminished by adopting a lifetime perspective, however, since individuals ultimately either spend or give away all of their incomes. Hence the distributional impact of excise taxes depends critically on the income elasticities of demand for goods and services subject to high rates of excise taxation. Poterba (1991) analyses US petrol taxes from the standpoint of lifetime incidence, finding that petrol consumption rises more than proportionately with affluence over much of the range of total spending, suggesting that petrol taxes are progressive, albeit less so than income taxes.

Optimal Excise Taxation

Ramsey (1927) initiated the modern theory of optimal taxation with his analysis of excise taxation in a model with identical consumers, finding that, far from being uniform, optimal excise tax rates vary inversely with elasticities of demand for taxed goods. Ramsey's set-up restricts the government to raising a given amount of revenue exclusively with excise taxes, and the resulting optimal tax pattern reflects that the excess burden of a tax increases with its behavioural impact. Diamond (1975) generalized the Ramsey rule to settings with heterogeneous individuals, showing that the resulting modified optimal excise taxes reflect both efficiency (lower tax rates on elastically demanded goods) and distributional (higher tax rates on goods purchased by wealthy individuals) considerations. As noted by Corlett and Hague (1953–4), the government's inability to tax leisure is what prevents uniform excise taxes from being optimal in the Ramsey model; as a second-best correction, the optimal Ramsey tax configuration entails imposing heavier excise taxes on goods and services that are complementary with untaxed leisure.

Under what circumstances would a government with access to a full range of income tax instruments want to impose excise taxes at differentiated rates? Atkinson and Stiglitz (1976) showed that, if consumers have identical utility functions that are weakly separable in consumption and leisure, then there is nothing to be gained by supplementing an optimal nonlinear income tax with differentiated excise taxes. The reason is that, in such a setting, patterns of commodity consumption fail to convey information to the government that is not already captured by income levels.

Excise taxes can nevertheless serve the function of controlling externalities, a consideration omitted from the Atkinson–Stiglitz framework. Pigou (1920) famously proposed the imposition of corrective excise taxes at rates equal to

marginal external damages, noting that doing so restores economic efficiency, and Sandmo (1975) illustrates the optimal application of Pigouvian excise taxes when the government relies on excise taxes to raise revenue. In practice, governments impose heavy taxes on energy products, motor vehicles and other transport, waste management, ozone- depleting substances, and other products and activities that arguably create externalities in degrading the environment. In 2000, OECD countries raised an average of 5.5 per cent of their total tax revenues from these environmental taxes, with European Union members averaging 6.8 per cent, and the United States the lowest in the OECD at 3.4 per cent (Hines 2007).

Excise taxes can also play a role in discouraging consumption of goods that may not have external effects, but are nonetheless harmful to the individuals who consume them. Examples of such goods include tobacco products, alcohol, and food with poor nutritional content. Irrational consumers may begin consuming these items without fully appreciating the regret they will experience years later, in which case there could be a role for optimal excise taxation to help consumers by making consumption more expensive, and therefore reducing the likelihood of consumers starting early on the path of over-consumption (O'Donoghue and Rabin 2006).

Finally, excise taxes raise enforcement concerns, as do all taxes. In the United Kingdom, which boasts the highest cigarette taxes in Europe, one cigarette out of every five is purchased on the black market (Cnossen and Smart 2005). Hence the choice of excise tax policy needs to be sensitive to smuggling and other evasion opportunities.

See Also

- ▶ Consumption Taxation
- ▶ Excess Burden of Taxation
- ▶ Optimal Taxation
- ▶ Pigouvian Taxes
- ▶ Progressive and Regressive Taxation
- ▶ Tax Compliance and Tax Evasion
- ▶ Tax Incidence
- ▶ Value-Added Tax

Bibliography

Atkinson, A.B.., and J.E. Stiglitz. 1976. The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics* 6: 55–75.

Besley, T.J., and H.S. Rosen. 1999. Sales taxes and prices: An empirical analysis. *National Tax Journal* 52: 157–178.

Cnossen, S. 1977. *Excise systems*. Baltimore: Johns Hopkins University Press.

Cnossen, S., and M. Smart. 2005. Taxation of tobacco. In *Theory and practice of excise taxation*, ed. S. Cnossen. Oxford: Oxford University Press.

Corlett, W.J., and D.C. Hague. 1953–4. Complementarity and the excess burden of taxation. *Review of Economic Studies* 21: 21–30.

Delipalla, S., and M. Keen. 1992. The comparison between ad valorem and specific taxation under imperfect competition. *Journal of Public Economics* 49: 351–368.

Diamond, P.A. 1975. A many-person Ramsey rule. *Journal of Public Economics* 4: 335–342.

Hines, J.R. Jr. 2007. Taxing consumption and other sins. *Journal of Economic Perspectives* 21(1): 49–68.

Hotelling, H. 1932. Edgeworth's taxation paradox and the nature of supply and demand functions. *Journal of Political Economy* 40: 577–616.

O'Donoghue, T., and M. Rabin. 2006. Optimal sin taxes. *Journal of Public Economics* 90: 1825–1849.

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Poterba, J.M. 1991. Is the gasoline tax regressive? In *Tax policy and the economy*, ed. D.F. Bradford, vol. 5. Cambridge, MA: MIT Press.

Poterba, J.M. 1996. Retail price reactions to changes in state and local sales taxes. *National Tax Journal* 49: 165–176.

Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.

Sandmo, A. 1975. Optimal taxation in the presence of externalities. *Swedish Journal of Economics* 77: 86–98.

Suits, D.B., and R.A. Musgrave. 1953. Ad valorem and unit taxes compared. *Quarterly Journal of Economics* 67: 598–604.

Executive Compensation

Michael L. Bognanno

Abstract

Chief executive officer (CEO) compensation is defined as the sum of base pay, bonuses, stock grants, stock options, other forms of

compensation and benefits. Inflation-adjusted, median total CEO compensation in the United States almost tripled between 1992 and 2000, with grants of stock options evolving to be the largest component of compensation. This article presents the arguments for and against this level and composition of CEO compensation.

Keywords

Agency framework; CEO; Executive compensation; Stock options

The level of executive pay has long been a flashpoint with the general public, particularly in periods of macroeconomic or stock market distress. Chief executive officer (CEO) compensation is defined as the sum of base pay, bonuses, stock grants, stock options, other forms of compensation and benefits. Of these components, grants of stock options have evolved to be the largest component of compensation. The near tripling of inflation-adjusted, median total CEO compensation between 1992 and 2000 was produced largely by a fivefold increase in stock options (Murphy 2002). To quantify the relative growth in CEO pay, in 1970 the average S&P 500 CEO earned approximately 30 times the pay of an average production worker. By 2002, this multiple had risen to almost 90 times the earnings of an average production worker in terms of CEO cash compensation (salary and bonus), and exceeded 360 times the earnings of an average production worker in terms of CEO total compensation (cash compensation, stock options and grants, and other compensation) (Hall and Murphy 2003).

In a comparison of 12 OECD countries, Abowd and Bognanno (1995) found that US CEOs were the most highly compensated. A similar result was not found for high-level US managers (human resource directors) or for US manufacturing operatives. (See also Murphy 1999 for international comparisons and a summary of theoretical and empirical research on executive pay.) Conyon and Murphy (2000) found that, controlling for firm size and other factors, US CEOs earned 190 per cent more in total compensation than UK CEOs. The United Kingdom's highest paid executive ranked only 97th on the list of the most highly paid US CEOs. Remarkably, Disney's Michael Eisner exercised options in 1997 worth more than the aggregate compensation of the top 500 UK CEOs.

While there has been an undisputed escalation in US CEO pay in absolute terms and in relation to the earnings of production workers over recent decades, academic researchers have taken positions on both sides of the debate over whether the level of CEO pay is economically justified or is the product of managerial power over the process of CEO pay determination. This article characterizes the basic framework within which CEO pay is viewed, and presents the principal arguments characterizing this debate.

The compensation of CEOs is generally viewed within an agency framework. (The gist of agency theory is found in the Bible, John 10:12: 'He who is a hired hand, and not a shepherd, who is *not the owner* of the sheep, sees the wolf coming, and leaves the sheep and flees, and the wolf snatches them and scatters them.') The separation that exists between the owners and managers in corporations gives rise to an agency problem in which managers have an incentive to pursue their personal interests over the interests of shareholders. The increase in the components of pay that are linked to firm performance, stock option schemes for example, are viewed from this perspective as a mechanism to align the incentives of managers with those of shareholders. Another incentive argument is based in tournament theory. The idea is that high CEO pay levels may increase the effort of those executives below the CEO position who are competing for promotion to the top spot (Lazear and Rosen 1981). However, reasons for caution against the use of intense competition also exist because competition may impede teamwork and spark counterproductive efforts (Lazear 1989).

Jensen and Murphy (1990) and Dow and Raposo (2005) attribute the sharp gain in CEO pay to the adoption of high-powered incentives in compensation packages. Hall and Liebman (1998) document the increasing responsiveness of CEO compensation to firm performance in the period between 1980 and 1994, a sensitivity due almost entirely to holdings of stock and stock

options. The argument has been made that the public focus on the level of CEO pay is misplaced because such concern changes the focus from the more important issue of how CEOs are paid and the link between CEO pay and firm performance (Jensen and Murphy 1990).

While the agency framework has provided a justification for stronger pay for performance and higher levels of contingent compensation, to some extent executives have responded to contingent compensation by seeking to avoid the risk created. A working paper by Ozerturk (2006) touches on the growth in the market for managerial hedging instruments, a growth coinciding with the growth in stockbased compensation in the late 1990s. The extent to which executives are reducing the pay–firm performance sensitivity of their compensation through the use of managerial hedging instruments is unknown, as disclosure rules are loose and participants in the market have no interest in publicizing their actions (Ozerturk 2006). Lavalle (2001) notes that at least 31 company insiders reported hedging in 2000, and for the majority it was a good decision. Hedging removes the agency theory basis for awarding large stock-based compensation packages.

The potential drawbacks of stock options are reviewed in Holden (2005), and include incentives for excessive risk taking and a focus on short-term performance, discouraging the payment of dividends and costing more to the firm than they are valued by risk-averse executives. Another negative feature of performance-based pay is the incentive created to manipulate or misstate the firm's financial performance. Efendi et al. (2007) find that the likelihood of accounting misstatements and severe accounting irregularities increases with the worth of CEO stock options.

The principal argument for the contention that the level of CEO pay is economically justified is that a competitive market for executive talent exists, and the level of CEO pay is a reflection of the intensive bidding by firms for scarce top talent. Tervio (2003) presents a competitive assignment model that determines CEO pay as the outcome of a bidding process between heterogeneous firms. In Gabaix and Landier's (2008) framework, CEOs of varying talent are matched to firms competitively, resulting in the largest firms having the top talent and the largest firms in the largest economies paying the most for CEO talent. Firm scale magnifies the pay discrepancies that result from small differences in CEO talent. High pay for top achievers is seen as appropriate given the value of their talent when magnified through the scale of large corporations. Their model also offers a partial explanation of the international differences in CEO pay.

Murphy and Zabojnik (2004) argue that CEO pay is determined by competitive forces, and the increase in CEO pay was driven primarily by an increase in the importance of general skills in running the modern corporation, as opposed to skills that are not transferable between firms, and the trend towards employing more externally hired CEOs. In the period from the 1970s to the 1990s, they state that the percentage of externally hired US CEOs rose from 15 per cent to more than 26 per cent. Others have also suggested that the trend toward filling the CEO position with external hires has played a role in increasing CEO pay.

In the presence of firms seeking to hire experienced CEOs from the outside labour market, Giannetti (2009) develops a model that offers an explanation of many aspects of managerial compensation, including benchmarking pay against larger firms, providing unrestricted stock awards to highly paid top executives and the use of long-term incentives. His model also offers an explanation of the growth in US CEO pay, and for international differences in CEO pay.

Further explanations arguing that the growth in CEO pay is justified also exist. Research suggests that CEO pay is in accord with historical norms in relation to the size of the firm, and that the marked growth in CEO pay is commensurate with growth in firm size. The link between firm size and CEO pay is very well documented (Murphy 1999). A recent study by Gabaix and Landier (2008) finds that the sixfold increase in CEO pay between 1980 and 2003 can be attributed to the sixfold increase in market capitalization of large US companies during that period.

A contrasting result is found in Bebchuk and Grinstein (2005). They find the growth in CEO

pay between 1993 and 2003 to exceed the increase than can be explained through changes in firm size, performance and industry mix. Mean compensation would have been only half as high in 2003 were the relationships between these factors and pay the same as they were in 1993.

The principal argument for the contention that CEOs are overpaid is the managerial power hypothesis. It argues that the pay-setting process is unduly influenced by the CEO, as the CEO may have substantial influence over the composition of the board of directors and of the compensation committee determining CEO pay (Crystal 1991). Support for this hypothesis is found in that CEO pay is excessive for firms with relatively weak boards of directors, for firms with no dominant outside shareholder, and for firms with a manager who has a relatively large ownership stake (Bebchuk et al. 2002; see Bebchuk and Fried 2003 for a review of the managerial power hypothesis).

The managerial power hypothesis also purports to explain several features of executive compensation, such as why stock option grants may have 'reload' provisions, why executives are allowed to exercise options early and hedge the risk of their options and stock holdings, and why US CEOs receive pay in excess of their international counterparts (Bebchuk et al. 2002). Murphy (2002) provides counter-arguments to the managerial power hypothesis.

CEO pay is both meaningful and growing in relation to firm financials. The ratio of the aggregate compensation paid by public companies to their top five executives to the aggregate earnings of their firms increased from five per cent in 1993–95 to 9.8 per cent in 2001–03 (Bebchuk and Grinstein 2005). This trend makes the debate over CEO pay levels and the policies directed towards CEO pay more important than ever. However, in pondering the regulation of CEO pay, there is a large literature with contrasting viewpoints that demand careful examination.

See Also

► Wage Inequality, Changes in

Bibliography

Abowd, J., and M. Bognanno. 1995. International differences in executive and managerial compensation. In *Differences and changes in wage structures*, ed. R. Freeman and L. Katz. Chicago: National Bureau of Economics Research (NBER)/University of Chicago Press.

Bebchuk, L.A., and J. Fried. 2003. Executive compensation as an agency problem. *Journal of Economic Perspectives* 17: 71–92.

Bebchuk, L.A., and Y. Grinstein. 2005. The growth of executive pay. *Oxford Review of Economics Policy* 21: 283–303.

Bebchuk, L.A., J. Fried, and D. Walker. 2002. Managerial power and rent extraction in the design of executive compensation. *University of Chicago Law Review* 69: 751–846.

Conyon, M., and K.J. Murphy. 2000. The prince and the pauper? CEO pay in the United States and the United Kingdom. *Economic Journal* 110: 640–671.

Crystal, G. 1991. *In search of excess: The overcompensation of the American executive*. New York: Norton.

Dow, J., and C. Raposo. 2005. CEO compensation, change, and corporate strategy. *Journal of Finance* 60: 2701–2727.

Efendi, J., A. Srivastava, and E. Swanson. 2007. Why do corporate managers misstate financial statements? The role of option compensation and other factors. *Journal of Financial Economics* 85: 667–708.

Gabaix, X., and A. Landier. 2008. Why has CEO pay increased so much? *Quarterly Journal of Economics* 123: 49–100.

Giannetti, M. 2009. Serial CEO incentives and the shape of managerial contracts. ECGI – Finance Working Paper No. 183/2007.

Hall, B., and J. Liebman. 1998. Are CEOs really paid like bureaucrats? *Quarterly Journal of Economics* 113: 653–691.

Hall, B.J., and K.J. Murphy. 2003. The trouble with stock options. *Journal of Economic Perspectives* 17: 49–70.

Holden, R. 2005. The original management incentive schemes. *Journal of Economic Perspectives* 19: 135–144.

Jensen, M.C., and K.J. Murphy. 1990. CEO incentives: It's not how much you pay, but how. *Harvard Business Review* 3: 138–149.

Lavalle, L. 2001. Commentary: Undermining pay for performance. *Business Week* 15: 70–71.

Lazear, E.P. 1989. Pay equality and industrial politics. *Journal of Political Economy* 97: 561–580.

Lazear, E.P., and S. Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.

Murphy, K.J. 1999. Executive compensation. In *Handbook of labor economics*, vol. 3, 1st ed, ed. O. Ashenfelter and D. Card, 2485–2563. Amsterdam: Elsevier.

Murphy, K.J. 2002. Explaining executive compensations: Managerial power versus the perceived cost of stock

options. *University of Chicago Law Review* 69: 847–869.
Murphy, K., and J. Zabojnik. 2004. CEO pay and appointments: A market-based explanation for recent trends. *American Economic Review* 94: 192–196.
Ozerturk, S. 2006. Hedge markets for executive and corporate agency. CORE Discussion Paper no. 2006/9.
Tervio, M. 2003. *The difference that CEOs make: An assignment model approach*. Mimeo: University of California, Berkeley.

Exhaustible Resources

Geoffrey Heal

Abstract

This article provides a brief overview of the ideas that have emerged in economics in connection with exhaustible resources. A resource is exhaustible if, the more we consume today, the less will be available for consumption at later dates. The dynamics of resource allocation, and the attainment of efficiency in intertemporal models, are thus key aspects of any economic theory of exhaustibility. The exhaustibility paradigm is widely applicable, including to climate change, biodiversity loss and even such non-environmental phenomena as antibiotic resistance.

Keywords

Biodiversity; Calculus of variations; Capital–resource ratio; Certainty-equivalence theorem; CES production function; Climate change; Cobb–Douglas function; Control theory; Depletion; Discount rate; Exhaustible resources; Hamiltonians; Hartwick rule; Hotelling model; Hotelling rule; Imperfect competition; Intertemporal substitutability; Neoclassical growth model; Resource rent; Shadow prices; Sustainability

JEL Classifications

Q3

Exhaustible resources are among the most important inputs to economic activities. Conventional crude oil and natural gas are good topical examples, with their pricing and availability currently a major source of concern. Of course, all minerals and extractive resources are exhaustible, as the volume of the earth is finite, but for many resources exhaustibility is not a matter of everyday concern as it is for oil and gas. There is a concern that the exhaustibility of low-cost deposits of oil and gas could restrict the long-run growth potential of the industrial world, and understanding the economic implications of exhaustibility is essential to grappling with this issue.

The availability of natural resources is a substantial topic in geology, with an extensive discussion there of the type of information available about resource supplies (see Barnett and Morse 1963; Brobst 1979; Smith 1980; Goeller 1979). Geologists classify resource stocks as proven, probable or possible reserves, which differ in the costs of extraction and the certainty with which their scale is known. An important lesson from the geological perspective is that the size of the total reserves of most resources is unknown. Even today, after many years of intensive prospecting, that remains true for a resource as important as oil.

The paradigm of exhaustibility is not limited in its applications to mineral resources: underground aquifers are exhaustible, and the capacity of the atmosphere to absorb greenhouse gases without radical change to the climate system has also been modelled as an exhaustible resource (Heal 1984). Another related and interesting exhaustible resource is the capacity to store carbon dioxide in underground rock formations: according to some perspectives on climate change, this could be a vitally important – and exhaustible – resource up to about 2050 (Butt et al. 1999). A very different recent application of this paradigm is in the field of drug resistance: Laxminarayan and Brown (2001) have modelled as an exhaustible resource the extent to which a drug can be used before resistance develops among the pathogens it is intended to kill. The world's stock of biodiversity can be seen as an exhaustible resource, too. Every time a species is driven to extinction, this stock falls in an irreversible way. We are depleting

that stock, and do not fully understand the consequences.

The first analytical discussion of exhaustibility can be found in the work of L.C. Gray in 1914, but the work that is regarded as the classic work in this area is that of Harold Hotelling in 1931. In an extraordinarily prescient article he provided the foundations of the modern theory of exhaustible resources, while simultaneously giving one of the earliest applications of calculus of variations in economics and developing an arbitrage-free model of equilibrium pricing. His work was so much ahead of its time that it was 30 years or more before it was fully appreciated. His name is now widely attached to the basic model of resource depletion, the 'Hotelling model', and to the movement of resource prices in a competitive market, which follow the 'Hotelling rule'.

A central feature of exhaustible resources is that they force us to think about resource allocation over time. For an exhaustible resource, the more we consume today, the less will be available for consumption at later dates. The dynamics of resource allocation, and the attainment of efficiency in intertemporal models, are key aspects of any economic theory of exhaustibility. This is probably the main reason why Hotelling's contribution was neglected for so long: in the 1930s economists were not ready to consider these issues analytically. It took the development of theories of growth, descriptive and optimal, to bring these to the fore.

The Hotelling Model

The simplest economic model of the use of an exhaustible resource is as follows. There is an initial stock S_0 of the resource, and we denote by $S_t \leq S_0$ the stock remaining at time t. Consumption of the resource at date t is c_t and the benefits from consumption are represented by a utility function $u(c_t)$, generally assumed to be strictly concave, twice differentiable and to satisfy a boundary condition such as $du/dc \to \infty$ as $c \to 0$ or $lim_{C \to 0}\ u(c) = -\infty$. Either of these conditions penalizes zero consumption very sharply and keeps consumption away from zero. Within this model we seek the time path of resource consumption that maximizes the present value of welfare:

$$\text{Max} \int_0^\infty u(c_t)\, e^{-\delta t} dt \text{ subject to } \int_0^\infty c_t dt \leq S_0 \tag{1}$$

Here $\delta \geq 0$ is a discount rate, and the second integral reflects the constraint imposed on cumulative use of the resource by its exhaustibility. It is this intertemporal conflict between present and future resource use that lies at the centre of theories of exhaustibility. Problem 1 is a classical problem in the calculus of variations, an isoperimetric problem, so called because it was first solved in the context of finding the closed plane curve having a given length and enclosing the greatest area. Economists have reformulated this problem so that it can be solved by methods from control theory rather than calculus of variations, and the reformulation usually used is

$$\text{Max} \int_0^\infty u(c_t)\, e^{-\delta t} dt \text{ subject to } \frac{dS_t}{dt} = -c_t, S_t \geq 0 \tag{2}$$

This replaces an integral constraint with a differential equation and a non- negativity constraint on a state variable. Solving this problem requires use of the Hamiltonian

$$H = u(c_t)e^{-\delta t} + \lambda_t e^{-\delta t}[-c_t]$$

where λ is a current value co-state variable. The first order conditions that a solution must satisfy are

$$u(c_t)' = \lambda_t \text{ and } \frac{1}{\lambda_t}\frac{d\lambda_t}{dt} = \delta \tag{3}$$

where a prime denotes the derivative of a function with respect to its argument, and two primes denote the second derivative. The first condition is intuitively obvious: the marginal utility of consumption must equal the shadow price of the resource. The second condition, which we call the Hotelling rule, states that the shadow price must rise at the discount rate. Another way of

looking at this is that in present value terms the shadow prices must be constant, so the present value of the marginal utility of consumption must be constant. Again, this is an intuitively satisfactory condition. It implies, of course, that consumption must fall over time.

To give an illustration of this result, consider the logarithmic utility function $u\ (c_t) = \ln c_t$. From 3 $\lambda_t = u(c_t)\ \prime = \frac{1}{c_t} = \lambda_0 e^{\delta t}$ so $c_t = c_0 e^{-\delta t}$ and $c_0 = \delta S_0$.

The Hotelling problem has also been known as the 'cake-eating problem' as it can be seen as choosing how to divide a finite cake between different periods. As long as the discount rate is positive there is a well-defined answer, as we have seen above. Matters are different if we try to treat all generations equally and set $\delta = 0$. Then, as we can see from the logarithmic example above, consumption is zero: indeed, it clearly goes to zero at all dates as the discount rate goes to zero. On an optimal path, as the discount rate falls we are trying to spread consumption ever more thinly over time and keep more for the future. This makes sense, except in the limiting case with a zero discount rate, in which it tells us never to consume anything and always to keep everything for the future. In this case problems 1 and 2 have no solution: although they look reasonable they are in fact ill-posed. There is an extensive literature on what exactly goes wrong in this case and what alternative formulations are available (Heal 1985; Gale 1967).

Extensions of the Basic Model

A simple change in the basic model is to allow for extraction costs, something that Hotelling did back in 1931. Consider a corporation extracting resources for a profit, rather than a national planning problem. The rate of extraction is again c_t and there is a cost of extraction of x per unit. The sale price of the resource is given by the demand curve $p(c_t)$, so the company is not a price-taker. The present value of profits is

$$\int_0^\infty [p(c_t)c_t - xc_t]e^{-\delta t}dt$$

and the corporation seeks to maximize this subject to the same constraints as above. The result is that

$$\left[p(c_t)'c_t + p - x\right] = \lambda_t \text{ and } \frac{1}{\lambda_t}\frac{d\lambda_t}{dt} = \delta$$

The first term is just the marginal revenue minus marginal cost: this difference again has to grow at the discount rate. We summarize this by saying that the rent to the resource must grow at the discount rate.

Another simple extension is to place a value on the remaining stock of the resource, $u(c_t, S_t)$. Think here of biodiversity in general or of a slow-growing forest. We value the flow of services that comes from depleting these, but we also value having some of the stock remaining (Krautkramer 1985; Heal 1998). Looking at the problem

$$\text{Max} \int_0^\infty u(c_t, S_t)e^{-\delta t}dt \text{ subject to } \frac{dS_t}{dt} = -c_t, S_t \geq 0 \quad (4)$$

leads to very different results. The first order conditions for optimality are that

$$u_c = \lambda \text{ and } \frac{d\lambda}{dt} - \delta\lambda = -u_s$$

where u_c, u_s are respectively the derivatives of u with respect to c_t and S_t. This solution is analysed at length in Heal (1998): what is interesting is that, provided we drop the boundary condition on $u(0)$ or $u'(0)$, there may be a stationary solution to this system of equations, at which $\frac{d\lambda}{dt} = 0, c_t = 0$ and $\delta = u_c/u_s$. This means that the discount rate equals the slope of an indifference curve in the $c - S$ plane, that is, it equals the marginal rate of substitution between the contributions to welfare of the stock and the flow.

Resources in Production

The natural next step in considering how exhaustible resources affect an economic system is to model how they enter into the production process and whether their exhaustibility acts as a drag on the economy, limiting its long-run growth.

E

This takes into the discussion of 'sustainability' and exhaustibility. The simplest way to do this is to take the basic neoclassical growth model and replace labour as an input by an exhaustible resource. This gives us as a production function $F(K_t, R_t)$ where K_t and R_t are respectively the capital stock and rate of resource use at date t. Typically F is assumed to be twice differentiable and to show constant or diminishing returns to the two inputs. With this formulation we can consider the problem

$$\text{Max} \int_0^\infty (c_t)e^{-\delta t}dt, c = F(K_t, R_t) - I_t, \quad \frac{dK_t}{dt} = I_t, \frac{dS_t}{dt} = -R_t, S_t \geq 0 \tag{5}$$

Here, obviously, I_t is investment at t, the rate of change of the capital stock. This problem is considerably more complex than any of the previous ones, and is analysed at length by Dasgupta and Heal (1974). Let $y_t = \frac{K_t}{R_t}, f(x_t) = F\left(\frac{K_t}{R_t}, 1\right)$, and $\eta(c) = \frac{-cu''(c)}{u'(c)}$ let σ be the elasticity of substitution between capital and the resource,

$$\sigma = \frac{-f'(x)f(x) - xf'(x)}{xf(x)f''(x)}$$

Then we can state the conditions characterizing a solution to Eq. (5) as follows (θ is the price of the produced good):

$$\frac{1}{\theta}\frac{d\theta}{dt} = F_K - \delta, \frac{1}{c}\frac{dc}{dt} = [F_K - \delta]/\eta(c) \tag{6}$$

$$\frac{1}{x}\frac{dx}{dt} = \sigma\frac{f(x)}{x} \text{ and } \lambda_t F_R e^{-\delta t} \text{ is constant} \tag{7}$$

Clearly in an economy with capital that can be reproduced and accumulated, and a resource that is exhaustible, growth must take place through the substitution of capital for resources. Equation (7) captures this process. It tells us that the capital–resource ratio changes at a rate that is the product of the elasticity of substitution and the average product per unit of fixed capital. The former indicates the ease with which substitution can be carried out, and the latter can be thought of as a measure of the importance of capital in production. So the easier it is to substitute, and the more important capital is, the more we substitute capital for resources.

What impact does the resource have on long-run growth possibilities? Clearly if $F(K, 0) > 0$ for $K > 0$ then the exhaustibility of the resource does not matter: we can continue producing when it runs out. The interesting case is $F(K, 0) = 0$ for any K. The fact that no production is possible without the resource does not mean that production must go to zero: consumption of the resource, as we saw in the Hotelling model, can be spread thinly over the indefinite future. If we can produce enough to support a good living standard with only a very small amount of the resource and a lot of capital, again exhaustibility may not matter. And, of course, technical progress may come to the rescue: it could increase the productivity of the scarce resource, or release the constraint that it imposes on production. In the light of these considerations Dasgupta and Heal (1979, p. 198) advance the following definitions: 'We shall regard an exhaustible resource as being *inessential* if there is a feasible program along which consumption is bounded away from zero: or in other words, if a positive sustainable level of consumption is feasible. Likewise, regard a resource as *essential* if feasible consumption must necessarily decline to zero in the long run.' This discussion anticipates many more recent discussions of sustainability. We can explore these issues further in the context of a CES production function, and in this case Dasgupta and Heal show that, if the elasticity of substitution between the resource and capital exceeds one, then the exhaustibility of the resource does not pose a fundamental problem. In the opposite case, an elasticity less than one, output must eventually fall to zero in the absence of technical progress. And in the borderline case, a unit elasticity, we have the Cobb–Douglas production function. In this case, if the elasticity of output with respect to capital exceeds that with respect to the resource, there is a feasible policy on which consumption is bounded away from zero. But in the remaining cases there is not: absent technical progress,

output must fall to zero. In fact, if we have a Cobb–Douglas production function with the elasticity of output with respect to capital greater than that with respect to the resource, not only are there paths on which output is bounded away from zero, but there are paths on which output is constant and on which output grows continuously. The paths on which output is constant may be characterized by a constant level of investment that maintains the total value of all stock constant, with the investment equal to the rent on the exhaustible resource. This is the Hartwick rule (Hartwick 1977; Asheim et al. 2003).

Backstop Technologies

Clearly the issue of substitutability is central to an understanding of the economic consequences of exhaustibility of important inputs. There are several dimensions to substitutability. One dimension, the one that we have discussed so far, is substitutability between capital and the resource. We can reduce oil use by insulating our buildings and wearing warmer clothes: this is substituting capital for oil. So is buying more expensive but more efficient furnaces. Both reduce oil consumption but require more capital. But there is another aspect of substitutability. Conventional crude oil is exhaustible and will be fully depleted at some point. But then there will be alternatives. For example, we can extract oil from coal – this is what Germany did during the Second World War and what South Africa did while it was subject to a trade embargo because of its apartheid policies. It is expensive by the standards of historical oil prices – perhaps $40 per barrel – but completely feasible. Similarly, oil can be extracted from tar sands – indeed, it is currently so extracted – but again at a cost that is high by historical oil price standards. And there are vast reserves of tar sands – they can probably provide more oil than all the conventional crude oil deposits in the Middle East. So other resources can replace oil when it runs out. This is a form of substitutability. Dasgupta and Heal (1974) modelled this by assuming that at a date T, which was unknown, the constraint imposed by the exhaustible resources would be lifted and an abundant substitute would become available. This is not unlike the situation described above with respect to oil and coal or tar sands. Another interpretation is that T is when nuclear fusion finally becomes a reality and, in the much-quoted phrase of the 1960s, energy finally becomes 'too cheap to meter'. (See also Nordhaus 1973, for simulation models of the effect of a backstop technology.)

To formalize this, assume that prior to T the technology is as in the previous section, but that after T there is a new technology that does not depend on the resource as an input. We can think of this as a dramatic technical change (such as fusion) or the appearance of an abundant substitute for the resource (such as tar sands for oil). Production from T onwards depends only on the capital stock available at T and not on the resource stock at that date. So we can write a state valuation function $V(K_T)$ giving the present value of welfare along an optimal policy from T onwards, discounted back to T. Then the overall problem is to

$$\text{Max} \int_0^T u(c_t)e^{-\delta t}dt + V(K_T), \text{ with for } t \in [0,T],$$
$$c = F(K_t, R_t) - I_t,\ \frac{dK_t}{dt} = I_t,\ \frac{dS_t}{dt} = -R_t, S_t \geq 0$$

Dasgupta and Heal assumed that T was unknown with density function ω_t. Then the maximand is the expected value of $\int_0^T u(c_t)e^{-\delta t}dt + V(K_T)$, which is

$$\int_0^\infty w_t \int_0^T \left[u(c_t)e^{-\delta t}dt + V(K_T)\right]$$

which on integration by parts and letting $\Omega_t = \int_t^\infty \omega_\gamma d\tau$ can be written as

$$\int_0^\infty e^{-\delta t}[u(c_t)\Omega_t + \omega_t V(K_t)]dt$$

Dasgupta and Heal (1974) explore possible solutions to this problem in some special cases, and characterize paths that are optimal. In the special case in which the valuation function V is independent not only of the resource stock ST (which we have assumed) but also of the capital stock K_T (so that the new source makes all existing capital obsolete) the optimum path satisfies

E

$$\frac{1}{x}\frac{dx}{dt} = \sigma\frac{f(x)}{x} \tag{8}$$

$$\frac{1}{c}\frac{dc}{dt} = \frac{F_K - \delta - \psi}{\eta(c)} \tag{9}$$

where $\psi_t = \frac{\omega_t}{\Omega_t}$. These are very similar to Eqs. (6) and (7) from the deterministic case with no backstop technology, except for the term Ψ that is added to the discount rate. Dasgupta and Heal also establish a certainty-equivalence theorem showing when the possibility of a backstop arriving can be subsumed entirely into a modification of the discount rate, by the addition of a term Ψ_t to the discount rate that reflects the conditional probability of the backstop arriving now, given that it has not yet arrived.

Of particular interest is the behaviour of the resource price when there is a possibility of a substitute arriving. A more direct way of understanding this is to look at a different model (Heal 1976), again with a backstop technology available as a replacement for the resource. Assume that there is a cost to extraction of the resource, and that this depends on and increases with the amount extracted to date. This is in many ways a natural assumption, in keeping what we know about the grade-tonnage distribution for most minerals. There are small amounts available at low extraction costs, more at larger costs and almost unlimited amounts if we are prepared to pay a sufficiently high cost. There is also a backstop available at a fixed unit cost, and in effectively unlimited supply.

One way of formalizing this is to let $z_t = \int_0^t c_\tau d\tau$ be the amount extracted to date, and have the current unit extraction cost depend on this:

$$\begin{aligned}&\text{extraction cost} = g(z),\\ &g' > 0 \text{ for } 0 \le z \le \overline{z} \text{ and}\\ &g(z) = \beta = g(\overline{z}) > 0 \text{ for } z \ge \overline{z}\end{aligned}$$

So the unit extraction cost is an increasing function of cumulative extraction up to a certain total extraction and a corresponding extraction cost, at which point a backstop becomes available in more or less unlimited amounts at a constant cost of β. In the case of a fixed extraction cost, which is the classic Hotelling model considered above, the price rises away from the marginal extraction cost at the discount rate – the rent on the resource satisfied Hotelling's rule and increases exponentially. This is natural: the rent on the resource reflects its scarcity, and this scarcity is rising over time, so it is natural for the rent to rise too. In the present case we should expect a different outcome. Overall the resource is not scarce: high-cost sources, to which we move over time, are abundant. It is only low-cost sources that are scarce. During the depletion of these, scarcity reigns, but once they have gone there is no longer scarcity in the sense of a limited availability. Any amount is available at the right price. So the dynamics of scarcity are in effect reversed. The solution reflects this, and is found by piecing together the solutions to two distinct problems:

$$\begin{aligned}&\text{Max}\int_0^\infty u(c_t)e^{-\delta t}dt \text{ subject to } c_t + \frac{dK_t}{dt}\\ &= F(K_t, R_t) - g(z_t)R_t\end{aligned} \tag{10}$$

and

$$\begin{aligned}&\text{Max}\int_0^\infty u(c_t)e^{-\delta t}dt \text{ subject to } c_t + \frac{dK_t}{dt}\\ &= F(K_t, R_t) - \beta R_t\end{aligned} \tag{11}$$

Problem 10 reflects the constraints on the economy up to the time when the backstop comes into use, and problem 11 represents the economy after all lower-cost reserves are depleted and once it is dependent on the backstop. We expect that a path that is optimal overall will first follow a solution to 10 and then one to 11, and this intuition can be verified formally (Heal 1976). Clearly, once we are in the second regime, we expect that the price of the resource will equal its marginal extraction cost β. What is less clear is how price and extraction cost are related during the phase corresponding to the solution to 10. It is possible to show the following: if λ is the price of the resource and θ that of the produced good, then

$$\frac{1}{\lambda}\frac{d\lambda}{dt} = \delta\left(1 - \frac{\theta g(z)}{\lambda}\right) + \frac{\theta g(z)}{\lambda}\left(\frac{1}{\theta}\frac{d\theta}{dt}\right) \tag{12}$$

Although it looks complicated, Eq. (12) in fact bears a simple and intuitive interpretation. It expresses the rate of change of the price of the resource as the weighted average of two terms, where the weights are $\left(1-\frac{\theta g(z)}{\lambda}\right)$ and $\frac{\theta g(z)}{\lambda}$, which respectively the fraction of the price that is pure rent and the fraction that is extraction cost. The two terms whose weighted average equals the rate of change of the resource price are the discount rate and the rate of change of the price of the produced good. So, if most of the price of the resource is rent, then its price rises at close to the discount rate, whereas if most of the price is extraction costs, then it rises at the rate at which extraction costs rise. This suggests a path of price movements rather different from the classic Hotelling model: in the present case prices will contain a large rent element early in the life cycle of the resource and then no rent at all towards the end of the life cycle.

The relationship between price and extraction cost, and so the movement of the scarcity rent over time, has been the subject of many additional papers, including Solow and Wan (1976), Hanson (1980), Farzin (1992) and Oren and Powell (1985). Oren and Powell extend the basic model presented above to consider a class of related issues, and Farzin focuses on the issue of whether the movement of scarcity rent must be monotonic. He concludes that there are reasonable cases in which the scarcity rent moves non-monotonically over time.

Imperfect Competition and Resource Use

So far we have considered patterns of resource use that are socially optimal, which are also the patterns of resource use that would emerge from a set of complete competitive markets with no elements of market failure present. While the answers to these questions are interesting and informative, it is clearly necessary to understand the impact of market imperfections on this picture. There is, not surprisingly, an extensive and sophisticated literature on this. I have space for only one or two basic insights.

We expect that moving from competition to monopoly will raise the prices of a good. To the extent that this is true for exhaustible resources, then monopoly will reduce the rate of extraction. Hence Robert Solow's comment that 'The monopolist is the conservationist's best friend'. But, with a fixed stock to sell, as in the basic Hotelling model, if the monopolist sells less now because of a higher price, then he must sell more in the future, when the benefits are discounted. This is unattractive to him: the intertemporal substitutability inherent in any choices about a time pattern of resource use brings a new dimension to the impact of monopoly or imperfect competition on price and output.

An interesting illustration of this is seen clearly in the simplest possible case, that of a monopolistic supplier of a resource with a zero marginal extraction cost. In this case, whether the monopolist charges a higher or a lower price initially than the competitive price depends on the behaviour of the price elasticity of demand for the resource along the demand curve. Consider the family of constant-elasticity demand curves, $p(c) = Ac^{-B}$ where $1 > B > 0$. The monopolist's problem is to

$$\text{Max} \int_0^\infty p(c)ce^{-\delta t}dt$$

subject to the usual constraint on the total availability of the resource. This requires that the marginal revenue from sale of the resource grows over time at the discount rate. Now, letting $\varepsilon(c)$ be the demand elasticity when consumption is c we can write marginal revenue as

$$MR_t = p_t(1 + 1/\varepsilon(c_t)) = p_t\gamma_t \text{ where } \gamma = (1 + 1/\varepsilon(c_t))$$

Hence

$$\frac{d\ln MR}{dt} = \delta = \frac{d\ln p}{dt} + \frac{dln\gamma}{dt}$$

and so

$$\frac{d\ln p}{dt} = \delta - \frac{dln\gamma}{dt}$$

For a constant elasticity demand curve it is easy to check that $\frac{dln\gamma}{dt} = 0$ so that in this case a

monopolist will want the price to rise at the discount rate – just as in the competitive case. More generally, we can show that

$$\text{If } \frac{d\eta}{dc} > (<)0 \text{ then } \frac{d\ln p}{dt} < (>)\delta$$

so that the nature of the bias from first best introduced by monopoly depends on the way the demand elasticity changes along the demand curve (Dasgupta and Heal 1979; Stiglitz 1976).

The case of a monopoly supplier is the simplest entry point to imperfect competition, but does not do justice to the sophistication of the results that are available in this area. One of the most interesting developments was motivated by the role of the OPEC cartel in the oil market, and a desire to understand its real long-run impact. This is the development of models of a market with a cartel and a 'competitive fringe', which seems to describe accurately the relationship between OPEC and non-OPEC members. Closely associated with this is the idea of limit pricing to keep a backstop technology out of the market. Models incorporating these ideas are summarized in Dasgupta and Heal (1979), and some of the key original articles are by Sweeney (1977), Gilbert (1978) and Pindyck (1978).

Conclusions

Exhaustible resources are economically important. In addition, exhaustibility is an analytically interesting property: it forces us to think even in the very simplest case about intertemporal issues, about the present versus the future. Without this conflict there is no exhaustible resource. So dynamics are integral in even the most basic thinking here, which is one reason why serious discussion of exhaustibility was so slow to emerge. It is not surprising that exhaustibility has featured centrally in discussions of sustainability, and earlier in the neo-Malthusian diatribes of the Club of Rome. Many of the issues that have emerged in the debates about sustainability have in fact been analysed by economists in the discussions of exhaustibility in the 1970s (see Heal 2003), usually with an emphasis on substitutability and technical progress as long-run solutions to the constraints imposed by exhaustibility, solutions that are typically more apparent to economists than to most others, though no less realistic for that. As I emphasized in the introduction, an interesting aspect of exhaustibility is the rather widespread applicability of the paradigm – to climate change, biodiversity loss and even such totally non-environmental phenomena as antibiotic resistance.

In this article I have been able to review only a fraction of a large and original literature on exhaustibility. I have certainly short-changed the literature on imperfect competition in markets for exhaustible resources, and have not touched at all on work on the empirical testing of the models of price movements discussed here (Heal and Barrow 1980; Miller and Upton 1985; Slade 1982; Agbeyegbe 1989). Another big gap is the theory of non-optimal growth with exhaustible resources, which asks how a market economy with imperfect futures markets will evolve. Because of the intrinsically intertemporal nature of the allocation problem with exhaustibility, the absence of a complete set of futures markets has particularly serious consequences. Again, there is an interesting and original literature on this (for a review, see Dasgupta and Heal 1979).

See Also

- ▶ Arbitrage Pricing Theory
- ▶ Hotelling, Harold (1895–1973)
- ▶ Miller, Merton (1923–2000)
- ▶ Pontryagin's Principle of Optimality
- ▶ Solow, Robert (Born 1924)
- ▶ Stiglitz, Joseph E. (Born 1943)

Bibliography

Agbeyegbe, T. 1989. Interest rates and metal price movements. *Journal of Environmental Economics and Management* 16: 184–192.

Asheim, G., W. Buchholz, and C. Withagen. 2003. The Hartwick rule: Myths and facts. *Environment and Resource Economics* 25: 129–150.

Barnett, H., and C. Morse. 1963. *Scarcity and growth*. Baltimore: Johns Hopkins University Press.

Brobst, D. 1979. Fundamental concepts for the analysis of resource availability. In *Scarcity and growth reconsidered*, ed. V. Kerry Smith. Baltimore: Johns Hopkins Press for Resources for the Future.

Butt, D., K. Lackner, C. Wendt, K. Nomura, and Y. Yanagisawa. 1999. The importance of and a method for disposing of carbon dioxide in a thermodynamically stable form. *World Resource Review* 11: 196–219.

Dasgupta, P., and G. Heal. 1974. The optimal depletion of exhaustible resources. In *Review of economic studies: Symposium on the economics of exhaustible resources*, 3–28.

Dasgupta, P., and G. Heal. 1979. *Economic theory and exhaustible resources*. Cambridge: Cambridge University Press.

Farzin, H. 1992. The time path of scarcity rent in the theory of exhaustible resources. *Economic Journal* 102: 813–830.

Gale, D. 1967. On optimal development in a multi sector economy. *Review of Economic Studies* 34: 1–18.

Gilbert, R. 1978. Dominant firm pricing policy in a market for an exhaustible resource. *Bell Journal of Economics* 9: 385–395.

Goeller, H. 1979. The age of substitutability: A scientific appraisal of natural resource adequacy. In *Scarcity and growth reconsidered*, ed. V. Kerry Smith. Baltimore: Johns Hopkins University Press for Resources for the Future.

Gray, L. 1914. Rent under the assumption of exhaustibility. *Quarterly Journal of Economics* 28: 466–489.

Hanson, D. 1980. Increasing extraction costs and resource prices: Some further results. *Bell Journal of Economics* 11: 335–342.

Hartwick, J. 1977. Intergenerational equity and investing the rents from exhaustible resources. *American Economic Review* 66: 9072–9074.

Heal, G. 1976. The relationship between price and extraction cost for a resource with a backstop technology. *Bell Journal of Economics* 7: 371–378.

Heal, G. 1984. Interactions between economy and climate: A framework for policy design under uncertainty. In *Advances in applied microeconomics*, ed. V. Kerry Smith and A. White, vol. 3. Greenwich/London: JAI Press.

Heal, G. 1985. Depletion and discounting: A classical issue in the economics of exhaustible resources. In *Environmental and natural resource mathematics*, *Proceedings of symposia in applied mathematics*, ed. R. McKelvey, vol. 32. Providence: American Mathematical Society.

Heal, G. 1998. *Valuing the future: Economic theory and sustainability.* New York: Columbia University Press.

Heal, G. 2003. Optimality or sustainability? In *Economics for an imperfect world: Essays in honor of Joseph Stiglitz*, ed. R. Arnott et al. Cambridge, MA: MIT Press.

Heal, G., and M. Barrow. 1980. The relationship between interest rates and metal price movements. *Review of Economic Studies* 47: 161–181.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39 (2): 137–175.

Krautkramer, J. 1985. Optimal growth, resource amenities and the preservation of natural environments. *Review of Economic Studies* 52: 153–170.

Laxminarayan, R., and G. Brown. 2001. Economics of antibiotics resistance: A theory of optimal use. *Journal of Environmental Economics and Management* 42: 183–206.

Miller, M., and C. Upton. 1985. A test of the Hotelling valuation principle. *Journal of Political Economy* 93: 1–25.

Nordhaus, W. 1973. The allocation of energy resources. *Brookings Papers on Economic Activity* 3: 529–576.

Oren, S., and S. Powell. 1985. Optimal supply of a depletable resource with a backstop technology: Heal's theorem revisited. *Operations Research* 33: 277–292.

Pindyck, R. 1978. Gains to producers from the cartelization of exhaustible resources. *Review of Economics and Statistics* 60: 238–251.

Slade, M. 1982. Trends in natural resource commodity prices: An analysis of the time domain. *Journal of Environmental Economics and Management* 9: 122–137.

Smith, V. Kerry. 1980. The evaluation of natural resource adequacy: Elusive quest or frontier of economic analysis? *Land Economics* 56: 257–298.

Solow, R., and F. Wan. 1976. Extraction costs in the theory of exhaustible resources. *Bell Journal of Economics* 7: 359–370.

Stiglitz, J. 1976. Monopoly and the rate of extraction of exhaustible resources. *American Economic Review* 66: 655–661.

Sweeney, J. 1977. Economics of depletable resources: Market forces and intertemporal bias. *Review of Economic Studies* 44: 125–141.

Existence of General Equilibrium

Gerard Debreu

Abstract

This article summarizes the history of the attempts to prove the existence of general equilibrium from those of Wald and others in Vienna in the 1930s to those of von Neumann and Nash, and of the solutions provided by Arrow, Debreu and McKenzie in the 1950s and their subsequent development.

Keywords

Arrow, K.; Convexity; Existence of general equilibrium; General equilibrium; Nash, J.; Neumann, J. von; Wald, A.; Walras, L.; Walras's Law

JEL Classifications

D5

Léon Walras provided in his *Eléments d'économie politique pure* (1874–7) an answer to an outstanding scientific question raised by several of his predecessors. Notably, Adam Smith had asked in *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776) why a large number of agents motivated by self-interest and making independent decisions do not create social chaos in a private ownership economy. Smith himself had gained a deep insight into the impersonal coordination of those decisions by markets for commodities. Only a mathematical model, however, could take into full account the interdependence of the variables involved. In constructing such a model Walras founded the theory of general economic equilibrium.

Walras and his successors were aware that his theory would be vacuous in the absence of an argument supporting the existence of its central concept. But for more than half a century that argument went no further than counting equations and unknowns and finding them to be equal in number. Yet for a non-linear system this equality does not prove that there is a solution. Nor would it provide a proof even for a linear system, especially when some of the unknowns are not allowed to take arbitrary real values.

A successful attack on the problem of existence of a general equilibrium was made possible by an exceptional conjunction of circumstances in Vienna in the early 1930s. It started from the formulation of the Walrasian model in terms of demand functions which had been given by Gustav Cassel in 1918. As Hans Neisser (1932) noted, certain values of commodity quantities and prices appearing in the solutions of Cassel's system of equations might be negative in such a way as to render those solutions meaningless. Heinrich von Stackelberg (1933) also made a cogent remark. Let x_i be the quantity of the ith final good demanded by consumers, a_{ij} the fixed technical coefficient specifying the input of the jth primary resource required for a unit output of the ith final good, and r_j the available quantity of the jth primary resource. The equality of demand and supply for every resource is expressed by

$$\sum_i a_{ij}x_i = r_j \text{ for all } j.$$

Von Stackelberg observed that if there are fewer final goods than primary resources, the preceding linear system of equations in $(x_1, \ldots, x_m)$ has, in general, no solution. Karl Schlesinger (1933–4) then remarked that equalities should be replaced by inequalities

$$\sum_i a_{ij}x_i \leqq r_j \text{ for all } j.$$

with the condition that a resource for which the strict inequality holds has a zero price. This suggestion, which had already been hinted at by Frederik Zeuthen (1932) in a different context, was essential to the proper formulation of the existence problem.

The problem thus posed received its first solution from Abraham Wald (1933–4), whose work on the existence of a general equilibrium gave rise to three published articles. The first two appeared in *Ergebnisse eines mathematischen Kolloquiums* in 1933–4 and in 1934–5. The third appeared in *Zeitschrift für Nationalökonomie* (1936) and was translated into English in *Econometrica* (1951). In that body of work Wald separately studied a model of production and a model of exchange and proved the existence of an equilibrium for each one.

By the standards prevailing in economic theory at that time, his mathematical arguments were of great complexity, and the major contribution that he had made did not attract the attention of the economics profession. A two-decade pause followed, and when research on the existence problem started again after 1950 it was under the

dominant influence of work done, also in the early 1930s, by John von Neumann. His article on the theory of growth, published in *Ergebnisse eines mathematischen Kolloquiums* (1935–6) and translated into English in the *Review of Economic Studies* (1945), contained in particular a lemma of critical importance. That lemma was reformulated in the following far more convenient form, and was also given a significantly simpler proof, by Shizuo Kakutani (1941). Let K be a non-empty, compact, convex set of finite dimension. Associate with every point x in K a non- empty, convex subset $\varphi(x)$ of K, and assume that the graph $G = \{(x, y) \in K \times K | y \in | \ \varphi(x) \}$ of the transformation φ is closed. Then φ has a fixed point x^*, i.e., a point x^* that belongs to its image $\varphi(x^*)$.

Kakutani's theorem was applied by John Nash, in a one-page note of 1950, to establish the existence of an equilibrium for a finite game. It can be used as well (Debreu 1952) to prove the existence of an equilibrium for a more general system composed of n agents. The ith agent chooses an action a_i in a set A_i of *a priori* possible actions. A state of the social system is therefore described by the list $a = (a_1, ..., a_n)$ of the actions chosen by the n agents. The preferences of the ith agent are represented by a real-valued utility function u_i defined for every a in the set of states $A = \times_{i=1}^{n} A_i$. Moreover the ith agent is restricted in the choice of his action in A_i, by the actions chosen by the other agents. Formally let N denote the set $\{1, \ldots, n\}$ of all the agents and N/i denote the set of the agents other than the ith. Let also $a_{N/i}$ denote the list of the actions $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ chosen by the agents in N/i. The ith agent is constrained to choose his own action in a subset $\varphi_i(a_{N/i})$ of A_i depending on $a_{N/i}$. In these conditions the ith agent, considering $a_{N/i}$ as given, chooses his action in $\mu_i(a_{N/i})$, the set of the elements of $\varphi_i(a_{N/i})$ at which the maximum of the utility function $u_i(\cdot, a_{N/i})$ in $\varphi_i(a_{N/i})$ is attained. Consider now the transformation $a \mapsto \mu(a) = \times_{i=1}^{n} \mu_i(a_{N/i})$ associating with any element a of A, the subset $\mu(a)$ of A. A state a^* is an equilibrium if and only if for every $i \in N$, the action a^*_i of the ith agent is best according to his preferences given the actions $a^*_{N/i}$ of the others, that is, if and only if for every $i \in N$, $a^*_i \in \mu_i(a^*_{N/i})$, that is, if and only if $a^* \in \mu(a^*)$. Thus the concept of an equilibrium for the social system is equivalent to the concept of a fixed point for the transformation $a \mapsto \mu(a)$ of elements of A into subsets of A. Ensuring that the assumptions of Kakutani's theorem are satisfied for the transformation μ yields a proof of existence of an equilibrium for the social system.

In the revival of interest in the problem of existence of a general economic equilibrium after 1950, the first solutions were published in 1954 by Kenneth Arrow and Gerard Debreu, and by Lionel McKenzie. The article by McKenzie emphasized international trade aspects, and the article by Arrow and Debreu dealt with an integrated model of production and consumption. Both rested their proofs on Kakutani's theorem. They were followed over the next three decades by a large number of publications (a bibliography is given in Debreu 1982) which confirmed the concept of a Kakutani fixed point as the most powerful mathematical tool for proofs of existence of a general equilibrium.

A simple prototype of the various economies that were the subject of those numerous existence results is (following Arrow–Debreu) composed of m consumers and n producers, producing, exchanging and consuming l commodities. The consumption of the ith consumer ($i = 1, \ldots, m$) is a vector x_i in R^l whose positive (or negative) components are his inputs (or outputs) of the l commodities. Similarly the production of the jth producer ($j = 1, \ldots, n$) is a vector y_j in R^l whose negative (or positive) components are his inputs (or outputs) of the l commodities. The ith consumer has three characteristics. (1) His consumption set X_i, a non-empty subset of R^l, is the set of his possible consumptions. (2) A binary relation $\precsim_i$ on X_i defines his preferences, and '$x_i \precsim_i x_i'$' is read as 'x_i' is at least as desired as x_i by the ith consumer'. Formally the preference relation of the ith consumer is the set $\{(x, x') \in X_i \times X_i | \ x \precsim_i x'\}$. (3) A vector e_i in R^l describes his initial endowment of commodities. The jth producer has one characteristic, his production set Y_j, a non-empty subset of R^l defining his possible productions. Finally the $\theta_{ij} \geq 0$ specifies the fraction of

E

the profit of the jth producer distributed to the ith consumer. These numbers satisfy the equality $\Sigma_{i=1}^{m}\theta_{ij} = 1$ for every j. In summary, the economy $\mathscr{E}$ is characterized by the list of mathematical objects

$$\left[(X_i, \precsim_i, e_i)_{i=1,\ldots,m}, (Y_j)_{j=1,\ldots,n}, (\theta_{ij})_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}\right].$$

Given a price-vector p in R^l different from 0, the jth producer ($j = 1, \ldots, n$) chooses a production y_j in Y_j that maximizes his profit, that is, such that the value $p \cdot y_j$ of y_j relative to p satisfies the inequality $p \cdot y_j \geq p \cdot y$ for every y in Y_j. Thus the ith consumer receives in addition to the value $p \cdot e_i$ of his endowment, $\Sigma_{j=1}^{n}\theta_{ij}\, p \cdot y_j$ the sum of his shares of the profits of the n producers. The value $p \cdot x$ of his consumption x is therefore constrained by the budget inequality $p \cdot x \leqq p \cdot e_i + \Sigma_{j=1}^{n}\theta_{ij}\, p \cdot y_j$. Under that constraint he chooses a consumption x_i in X_i that is best according to his preferences. The list $[p, (x_i)_{i=1,\ldots,m}, (y_j)_{j=1,\ldots,n}]$ of a non- zero price-vector, m consumptions and n productions forms a general equilibrium of the economy $\mathscr{E}$ if for every commodity, the excess of demand over supply vanishes,

$$\sum_{i=1}^{m} x_i - \sum_{j=1}^{n} y_j - \sum_{i=1}^{m} e_i = 0.$$

The existence of a general equilibrium can be proved (following Arrow–Debreu) by casting the economy $\mathscr{E}$ in the form of a social system of the type defined above. For this it suffices to introduce, in addition to the m consumers and to the n producers, a fictitious price-setting agent whose set of actions and whose utility function are now specified. Note first that the definition of a general equilibrium is invariant under multiplication of the price-vector p by a strictly positive real number. In the simple case where all prices are non-negative, one can therefore restrict p to be an element of the simplex $P = \left\{p \in R^l_+ \mid \Sigma^l_{h=1} p^h = 1\right\}$, the set of the vectors in R^l whose components are non-negative and add up to one. The set of actions of the price-setter is specified to be P. Given the consumptions $(x_i)_{i=1,\ldots,m}$ chosen by the m consumers, and the productions $(y_j)_{j=1,\ldots,n}$ chosen by the n producers, there results an excess demand

$$z = \sum_{i=1}^{m} x_i - \sum_{j=1}^{n} y_j - \sum_{i=1}^{m} e_i.$$

The utility function of the price-setter is specified to be $p \cdot z$. Maximizing the function $p \mapsto p \cdot z$ over P carries to one extreme the idea that the price-setter should choose high prices for the commodities that are in excess demand, and low prices for the commodities that are in excess supply.

Some of the assumptions on which the theorems of Arrow–Debreu (1954) are based are weak technical conditions: closedness of the consumption-sets, of the production-sets and of the preference relations, existence of a lower bound in every coordinate for each consumption-set, possibility of a null production for each producer. Other assumptions were later shown to be superfluous for economies with a finite set of agents: irreversibility of production (if both y and $-y$ are possible aggregate productions, then $y = 0$), free disposal (any aggregate production $y \leq 0$ is possible), and completeness and transitivity of preferences. Convexity of preferences can be dispensed with, and convexity of consumption-sets can be weakened, in economies with a large number of small consumers. Insatiability of consumers is an acceptable behavioural postulate. There remain, however, two overly strong assumptions. They are the hypothesis that for every i, the endowment e_i yields a possible consumption for the ith consumer (after disposal of a suitable commodity- vector if need be), and the assumption of convexity on the total production-set $Y = \Sigma_{j=1}^{n} Y_j$ which implies non-increasing returns to scale in the aggregate.

An alternative approach to the problem of existence of a general equilibrium, closer to traditional economic theory, is centred on the concept of excess demand function, or of excess demand correspondence. Given an economy $\mathscr{E}$ defined as before, consider a price-vector p in R^l_+ different from 0. The productions $(y_1, \ldots, y_n)$ chosen by the producers, and the consumptions $(x_1, \ldots, x_m)$ chosen by the consumers in reaction to the

price-vector p result in an excess demand z in the commodity-space R^l. If z is uniquely determined, the excess demand function f from $R^l_+/0$ to R^l is there by defined. If z is not uniquely determined, the set of excess demands in R^l associated with p is denoted by $\varphi(p)$, and the excess demand correspondence φ is thereby defined on $R^l_+/0$. Both f and φ are homogeneous of degree zero since $f(p)$ and $\varphi(p)$ are invariant under multiplication of p by a strictly positive real number. This permits various normalizations of p. For instance, p may be restricted to the simplex P. Moreover, for every $i = 1, ..., m$, one has $p \cdot x_i \leq p \cdot e_i + \Sigma_{j=1}^{n} \theta_{ij} p \cdot y_j$. By summation over i, one obtains

$$p \cdot \sum_{i=1}^{m} x_i \leq p \cdot \sum_{i=1}^{m} e_i + p \cdot \sum_{j=1}^{n} p.y_j,$$

or equivalently $p \cdot z \leq 0$. Therefore for every p in $R^l_+/0$, one has either $p \cdot f(p) \leq 0$ or $p \cdot \varphi(p) \leq 0$. This observation leads to the following proof of existence of a general equilibrium (Gale 1955; Nikaidô 1956; Debreu 1956). Let φ be a correspondence transforming points of the simplex P into non-empty convex subsets of R^l. If φ is bounded, has a closed graph and satisfies $p \cdot \varphi(p) \leq 0$ for every p in P, then, by Kakutani's theorem, there are a point p^* in P and a point z^* in R^l such that $z^* \in \varphi(p^*)$ and $z^* \leq 0$. In economic terms, there is a price-vector p^* in P yielding an associated excess demand z^* in $\varphi(p^*)$, all of whose components are negative or zero.

If all the consumers in the economy $\mathscr{E}$ are insatiable, every individual budget constraint is binding, and one has for every $i, p \cdot x_i = p \cdot e_i + \Sigma_{j=1}^{n} Y \theta_{ij} p \cdot y_j$. By summation over i, $p \cdot z = 0$. Thus in the case where the vector z associated with p is uniquely determined, the excess demand function satisfies

Walras's Law: for every p in $R^l_+/0$, $p \cdot f(p) = 0$.

In geometric terms, in the commodity-price space R^l the vectors p and $f(p)$ are orthogonal. This prompts one to normalize the price-vector p so that it belongs to the positive part of the unit sphere $\overline{S} = \{p \in R^l_+ | \, \|p\| = 1\}$, for then $f(p)$ can be represented as a vector tangent to $\overline{S}$ at p. The excess demand function is now seen as a vector field on $\overline{S}$. This in turn suggests another proof of existence of a general equilibrium (Dierker 1974) for the particular case of an exchange economy E whose consumers have continuous demand functions, monotone preferences and strictly positive endowments of all commodities. In that case for every $i = 1, ..., m$, the consumption-set X_i of the ith consumer is R^l_+ and $x < x'$ implies $x \prec_i x'$ (if x' is at least equal to x in every component and $x' \neq x$, then x' is preferred to x). Since the demand of a consumer with monotone preferences is not defined when some prices vanish, one must restrict the price-vector p to be strictly positive in every component, that is, to belong to $S = \{p \in \text{Interior } R^l_+ | \, \|p\| = 1\}$. Moreover let p_q be a sequence of price-vectors in S converging to p_0 in the boundary $\overline{S} / S$ of S. Thus for every q, the vector p_q is strictly positive in each component, while, in the limit, p_0 has some zero components. Then the associated sequence of excess demands $f(p_q)$ is unbounded. As a consequence, the vector field f points inward towards S near the boundary of S. In these conditions Brouwer's fixed point theorem yields the existence of an equilibrium price-vector p^* in S for which excess demand vanishes, $f(p^*) = 0$.

The preceding solutions of the problem of existence of a general equilibrium all rest directly on fixed point theorems. Three different lines of approach are provided by (1) combinatorial algorithms for the computation of approximate general equilibria, (2) differential processes converging to general equilibria, (3) the theory of the fixed point index of a mapping.

1. The past two decades have witnessed the development of algorithms of a combinatorial nature for the computation of an approximate general equilibrium (*see* computation of general equilibria). Given any number $\varepsilon > 0$, a constructive procedure thereby yields a price-vector p such that the norm $|f(p)|$ of the associated excess demand is smaller than ε. A compactness argument then gives a sequence of price-vectors p_q in S converging to p_0 for which $|f(p_q)|$ tends to 0. In the limit, $f(p_0) = 0$.
2. Global analysis was introduced into economic theory at the beginning of the 1970s to study

the set of general equilibria of an economy and the manner in which it depends on the economy. In that framework Stephen Smale proposed in (1976) a differential process which starts from a point in the boundary of the set of normalized price-vectors, and which converges to the set of equilibria provided that the initial point does not lie in a negligible exceptional set (*see* global analysis in economic literature). Another constructive procedure thus gives, from a differentiable viewpoint, conditions under which the set of general equilibria is not empty.

3. In the same differentiable framework Egbert Dierker (1972) used the theory of the fixed point index of a mapping to prove that a regular economy (as defined by him in regular economies below) whose excess demand points inward near the boundary of S has an odd (hence non-zero) number of general equilibria. The significance of this theorem rests on the fact that under its assumptions almost every economy is regular.

The previous existence results have been extended in many directions. The study of the core of an economy led to the consideration of a set of agents, all of whom are negligible relative to their totality. This concept was formalized first as an atomless measure space of agents, and later by means of non-standard analysis. In both cases the existence of a general equilibrium had to be proved for economies with infinitely many agents.

In order to specify a commodity one lists its physical characteristics, the date, the location, and the event at which it is available. As soon as one of those four variables can take infinitely many values, the analysis of general equilibrium must be set in the framework of infinite-dimensional commodity spaces. Several existence results were obtained in that context.

In yet another direction, external effects called for extensions. When the characteristics of each agent (e.g. his preferences, his production set, . . .) depend on the actions chosen by the other agents, formulating the economy as a social system of the type described earlier immediately yields an existence theorem. Still other extensions have covered economies with public goods, with indivisible commodities, and with non-convex production sets.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Fixed Point Theorems

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Arrow, K.J., and M.D. Intriligator, eds. 1981–6. *Handbook of mathematical economics*. Vol. I–III. Amsterdam: North-Holland.

Cassel, K.G. 1918. *Theoretische Sozialökonomie*. Leipzig: C.F. Winter.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences* 38: 886–893.

Debreu, G. 1956. Market equilibrium. *Proceedings of the National Academy of Sciences* 42: 876–878.

Debreu, G. 1982. Chapter 15: Existence of competitive equilibrium. In Arrow and Intriligator (1981–6).

Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40: 951–953.

Dierker, E. 1974. *Topological methods in Walrasian economics*. Berlin/New York: Springer-Verlag.

Gale, D. 1955. The law of supply and demand. *Mathematica Scandinavica* 3: 155–169.

Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8: 457–459.

McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.

Nash, J.F. 1950. Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences of the USA* 36: 48–49.

Neisser, H. 1932. Lohnhöhe und Beschäftigungsgrad im Marktgleichgewicht. *Weltwirtschaftliches Archiv* 36: 415–455.

Nikaidô, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8: 135–145.

Schlesinger, K. 1933–4. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 6: 10–20.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3: 107–120.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, vol. 2, ed. R.H. Campbell, A.S. Skinner, and W.B. Todd. Oxford: Clarendon Press, 1976.

von Neumann, J. 1935–36. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* 8: 73–83. Trans. by G. Morgenstern as 'A model of general economic equilibrium'. *Review of Economic Studies* 13(1), 1945, 1–9.

von Stackelberg, H. 1933. Zwei kritische Bemerkungen zur Preistheorie Gustav Cassels. *Zeitschrift für Nationalökonomie* 4: 456–472.

Wald, A. 1933–4. Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen Kolloquiums* 6: 12–20.

Wald, A. 1934–5. Über die Produktionsgleichungen der ökonomischen Wertlehre. *Ergebnisse eines mathematischen Kolloquiums* 7: 1–6.

Wald, A. 1936. Über einige Gleichungssysteme der mathematischen ökonomie. *Zeitschrift für Nationalökonomie* 7: 637–670. Trans. by Otto Eckstein as 'On some systems of equations of mathematical economics'. *Econometrica* 19(4), 1951, 368–403.

Walras, L. 1874–7. *Éléments d'économie politique pure*. Lausanne: L. Corbaz. Trans. by William Jaffé as *Elements of pure economics.* Homewood: Richard D. Irwin, 1954.

Zeuthen, F. 1932. Das Prinzip der Knappheit, technische Kombination und ökonomische Qualität. *Zeitschrift für Nationalökonomie* 4: 1–24.

Exit and Voice

Albert O. Hirschman

Abstract

Exit and voice are alternative responses to an unsatisfactory relationship: exit is the withdrawal from it, voice is the attempt to improve it through communication. They are not mutually exclusive responses: thus, the market is the archetypal exit mechanism, yet it usually involves voice. When available jointly, exit and voice may reinforce or undercut each other: the exit option enhances the influence of customers' voice on an unsatisfactory supplier but also reduces its volume. Exit–voice analysis has been applied to trade unions, hierarchies, public services, migration and political action, political party systems, marriage and divorce, and adolescent development.

Keywords

Adolescent development; Asymmetric information; City–suburb migration; Collective voice; Democratization; Education finance; Exit and voice; Free rider problem; Health finance; Hierarchy; Hirschman, A. O.; Horizontal vs vertical voice; Identity; Incomplete information; International capital flows; International migration; Loyalty; Marriage and divorce; Montesquieu, C. de; Multiparty systems; Political parties; Public services; Rural–urban migration; Smith, A.; Smith, A.; Trade unions; Two-party systems; Vouchers; Welfare state

JEL Classifications

B4

A central place is held in economics and social science in general by principles and forces making for order or equilibrium in economic and social systems. Disorder and disequilibrium are then understood as resulting from some malfunction of these principles or forces. Explanations of order–disorder or equilibrium–disequilibrium have typically been discipline-bound, dealing with either the political or the economic world. Since the two are interrelated it would be useful to have a construct that bridges them. Such is the claim of the exit–voice perspective. It addresses the changing balance of order and disorder in the social world by pointing out that social actors who experience developing disorder have available to them two activist reactions and perhaps remedies: *exit*, or withdrawal from a relationship that one has built up as a buyer of merchandise or as a member of an organization such as a firm, a family, a political party or a state; and *voice*, or the attempt at repairing and perhaps improving the relationship through an effort at communicating one's complaints, grievances and proposals for improvement. The voice reaction belongs in good part to the political domain since it has to do with the articulation and channelling of opinion, criticism and protest. Much of the exit reaction, on the contrary, involves the economic realm

as it is precisely the function of the markets for goods, services, and jobs to offer alternatives to consumers, buyers and employees who are for various reasons dissatisfied with their current transaction partners.

The exit–voice alternative was proposed and explored in *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States (Hirschman* 1970*, henceforth EVL).* Attempts to apply the book's perspective were made over many areas of social life. In the following, the basic concepts will be recapitulated and, where necessary, reformulated. Subsequently some major applications of the exit–voice polarity will be reviewed.

Basic Concepts

Exit

Exit means withdrawal from a relationship with a person or organization. If this relationship fulfils some vital function, then the withdrawal is possible only if the same relationship can be re-established with another person or organization. Exit is therefore often predicated on the availability of choice, *competition*, and well functioning *markets*.

Exit of customers (or employees) serves as a signal to the management of firms and organizations that something is amiss. A search for causes and remedies will then be undertaken and some plan of action designed to restore performance will be adopted. This is one way in which markets and competition work to prevent decay and to maintain and perhaps improve quality.

Exit is a powerful but indirect and somewhat blunt way of alerting management to its failings. Most of the time, those customers and members of organizations who exit have no interest in improving them by their withdrawal, so that exit does not provide management with much information on what is wrong.

Voice

The direct and more informative way of alerting management is to alert it: this is *voice*. Its role is, or should be, paramount in situations where exit is either not available at all or is difficult, costly, and traumatic. This is so for certain primordial groupings one is born into – the family, the ethnic or religious community, the nation – or for those organizations one joins with the intention of staying for a prolonged period – school, marriage, political party, firm. With regard to buying and selling, voice should take over from exit when competition is weak or nonexistent as in the case of goods and services being produced under oligopolistic or monopolistic conditions, or when exit is expensive for both parties as in certain interfirm relations.

Unlike exit in the case of well-functioning markets, voice is never easy. It can even be dangerous. Many organizations and their agents are not at all keen on being told about their shortcomings by members and the latter often expose themselves to reprisals if they utter any criticism (Birch 1975). Even in the absence of reprisals, the cost of voice to an individual member will often exceed, in terms of time and effort, any conceivable benefit from voicing. Frequently, moreover, any effective channelling of individual voices requires a number of members to join together so that voice formation depends on the potential for collective action.

In spite of these problems, voice exists or, rather, it has come into being. Its history is to a considerable extent the history of the right to dissent, of due process, of safeguards against reprisal, and of the advance of trade unions and of consumer and many other organizations articulating the demands of individuals and groups who once were silent. Similarly, the history of exit is the history of the broadening of the market, of the right to move freely, to emigrate, to be a conscientious objector, to divorce, etc. Being two basic, complementary ingredients of democratic freedom, the right to exit and the right to voice have on the whole been enlarged or restricted jointly. Yet, there are important instances of unilateral advances or retreats of either the one or the other response mechanism (Rokkan 1975; Finer 1974).

Interaction of Exit and Voice

As noted, exit is paramount as a reaction to discontent in some circumstances and voice holds a

similarly privileged position in others, but frequently both mechanisms are available jointly. In such situations they may either reinforce or undercut each other. The availability and threat of exit on the part of an important customer or group of members may powerfully reinforce their voice. On the other hand, the actual recourse to exit will often diminish the volume of voice that would otherwise be forthcoming and, should the organization be more sensitive to voice than to exit, the stage could be set for cumulative deterioration. For example, after an incipient deterioration of public schools or inner cities, the availability of private schools or suburban housing would lead, via exit, to further deterioration – a turn of events that might have been prevented if the parents sending their children to private school or the inner city residents who move to the suburbs had instead used their voice to press for reform. In their aggregate effects, the individual exit decisions are harmful – an instance of the 'tyranny of small decisions' – also because they are likely to be taken on the basis of a short-run private-interest calculus only and do not take into account the 'public bad' that will be inflicted, even on those who exit, by decaying inner cities and segregated education (Levin 1983; Breneman 1983).

These kinds of situations are sufficiently numerous and important to be of interest not only as curious paradoxes showing that under some circumstances the availability of exit (that is, of competition) could have undesirable effects. In this connection, *EVL* stressed the value of *loyalty* as a factor that might delay over-rapid exit. Loyalty would make a member reluctant to leave an organization upon the slightest manifestation of decline even though rival organizations were available. Provided it is not 'blind', loyalty would also activate voice as loyal members are strongly motivated to save 'their' organization once deterioration has passed some threshold.

The difficulties of combining exit and voice in an optimal manner are in a sense 'problems of the rich': they relate to situations and societies where exit and voice are both forthcoming more or less abundantly, but where, for best results, one would wish for a different mix. Historically more frequent are cases where exit and voice are both in short supply, in spite of many reasons for discontent and unhappiness. There is no doubt, as many commentators have pointed out, that passivity, acquiescence, inaction, withdrawal, and resignation have held sway much of the time over wide areas of the social world. This is largely the result of repression of both exit and voice – a repression that has flourished in spite of the fact that all human organizations could put to good use the feedback provided by the two reaction modes.

E

Problems in Voice Formation

The development of voice among customers of firms or members of organizations poses a number of problems that were not fully explored in *EVL*. Critics have asserted that, in its endeavour to present voice as a ready alternative to exit, the book understated the difficulties of voice formation. In examining this issue it is useful to start with the extreme no-voice case: the authoritarian state which is dedicated to repressing and suppressing voice. This situation has given rise to a useful distinction between *horizontal* and *vertical* voice (O'Donnell 1986). The latter is the actual communication, complaint, petition, or protest addressed to the authorities by a citizen and, more frequently, by an organization representing a group of citizens. Horizontal voice is the utterance and exchange of opinion, concern and criticism *among* citizens: in the more open societies it is today regularly ascertained through opinion polls revealing the approval rating of presidents, prime ministers, mayors, etc. Horizontal voice is a necessary precondition for the mobilization of vertical voice. It is the earmark of the more frightful authoritarian regimes that they suppress not only vertical voice – any ordinary tyranny does that – but horizontal voice as well. The suppression of horizontal voice is generally the side-effect of the terrorist methods used by such regimes in dealing with their enemies.

The distinction between vertical and horizontal voice is relevant to the 'free ride' argument in relation to voice formation (Barry 1974). For *vertical* voice to come about, that is, for members of the organization to engage management in

meaningful dialogue, it is frequently necessary for members to forge a tie among themselves, to create an organization which will agitate for their demands, etc. But the hoped-for result of collective voice is a freely available public good; hence, so goes the critical argument, self-interested, 'rational' individuals may well withhold their contribution to the voice enterprise in the expectation that others will take on the entire burden. Important as it is, this argument has its limitations. First of all, it is addressed only to vertical voice which it mistakenly equates (as *EVL* did) with voice in general. Horizontal voice is not subject to the strictures of the free-rider argument: it is free, spontaneous activity of men and women in society, akin to breathing. As just noted, extraordinary violence has to be deployed if it is to be suppressed. Under ordinary circumstances, horizontal voice is continuously generated and has an impact even without becoming vertical: in many environments managers of organizations cannot help noticing and reacting to critical opinions and hostile moods of the members, whether or not organized protest movements break out. That the planned economies of Eastern Europe function to the extent they do has been explained on precisely this ground (Bender 1981, p. 30).

Another limitation of the free-rider argument lies in its assumption that individuals will always act instrumentally. Just because the desired result of collective voice is typically a *public* good – or, better, some aspect of the *public* happiness – participation in voice provides an alternative to self-centred, instrumental action. It therefore has the powerful attractions of those activities that are characterized by the fusion of striving and attaining and can be understood as investments in individual or group identity (Hirschman 1985).

Some Areas of Application

Trade Unions

In economics, the major application of the exit–voice theme has been the analysis of trade unions as collective voice by Freeman and Medoff in their book *What Do Unions Do?* (1984). Instead of looking at unions as a monopolistic device raising wages for unionized workers beyond the 'market-clearing' equilibrium level or – much the same zero-sum interpretation in different language – as a tool in the class struggle serving to reduce the degree of exploitation, the book finds that a major function of unions is that of channelling information to management about workers' aspirations and complaints. Collective voice, in the form of union bargaining, is more efficient in conveying information about workers' discontent – and in doing something about it – than individual decisions to quit, as voice carries more information than exit. The presence of union voice is shown to reduce costly labour turnover. Moreover, the fringe benefits, workplace practices, and seniority rules which unions negotiate often result in offsetting labour productivity increases.

Markets and Hierarchies vs. Exit and Voice

Renewed attention has been given in recent years to the question why some kinds of economic activities are carried on through many independent firms while others, to the contrary, are tied together through bureaucratic and hierarchical relations. In accounting for hierarchy, one approach has directed attention to such matters as uncertainty about the evolution of the market and the technology and in particular to asymmetric availability of information to buyer and seller, creating opportunities for deceitful behaviour (Williamson 1975). Hierarchy is then seen as superior to markets whenever there is need for a sustained and frank dialogue between the contracting parties. Critics of this position have argued: (1) relations between independent firms, such as contractors and subcontractors, are often quite effective in discouraging malfeasance; (2) correlatively, hierarchy frequently leads to characteristic patterns of concealment and control evasion (Eccles 1981; Granovetter 1985); and (3) industry structure varies substantially from one country to another as well as within the same country over time: in Japan, for example, subcontracting is much more widely practised than in the West and in Italy subcontracting has become more widespread in the last 10–20 years.

A formulation in terms of exit–voice is helpful here. The characteristics which are said to justify hierarchy – incomplete information, considerable apprenticing of one firm by the other, openings for 'opportunistic' (i.e., dishonest) behaviour, etc. – all make for situations in which there is need for voice: the firms contracting together must intensively consult with, and watch over, each other. *But the need for voice does not necessarily imply that hierarchy is in order.* Whether voicing is done best within the same organization or from one independent firm to another is by no means a foregone conclusion. Moreover, when the two parties are independent and resort a great deal to voice, the possibility of exit from the relationship often looms in the background. The implicit threat of exit could carry as much clout as that of sanctions in hierarchical relationships.

The argument for hierarchy in cases where voice has an important role to play may arise from thinking of market relationship only in terms of the ideal, anonymous market where voice is wholly absent. But most markets involve voice: commerce *is* communication, and is premised on frequent and close contact of the contracting parties who deliver promises, trust them, and engage in mutual adjustment of claims and complaints – all of this was implicit in the eighteenth-century notion of *doux commerce* (Hirschman 1977, 1982). Adam Smith even conjectured that it was man's ability to communicate through speech that lies at the source of his 'propensity to truck and barter'. How odd, then, that the need for frequent and intensive communication should be adduced as a conclusive argument for hierarchy.

Public Services: Education, Health, Others

The organization of public services represents a privileged area for the application of exit–voice reasoning – significantly the exit–voice idea had its origin in the analysis of a public service in trouble, the Nigerian railroads (*EVL*, Preface). Public services are typically sold or delivered by a single public or publicly regulated supplier, for various well-known reasons.

With the production of most public services being thus deprived of the 'discipline of the market', problems of productive efficiency and quality maintenance arise necessarily. An obvious way of mitigating these problems is to attempt to reintroduce market pressures in some fashion. For example, when certain categories of goods and services are to be made available either to all citizens regardless of their income or to some deprived social groups, the state and its agencies can sometimes refrain from producing or distributing these goods directly, and instead issue special purpose money or *vouchers* enabling the beneficiaries to acquire the goods or services through ordinary market channels. In this manner the voucher system reintroduces the market and the possibility of exit. A particularly successful example of the voucher system is the distribution of Food Stamps to low-income persons in the United States. Instead of creating and administering its own food distribution network the state hands out vouchers (food stamps) which the beneficiaries can then use at existing, competitive commercial outlets.

In part because of the success of this programme and in part because of the belief in 'market solutions' as the remedy for all that ails government programmes, voucher schemes have been proposed for a large number of other public services, from education to low-cost housing to the supply of certain health services. Voucher systems are appropriate primarily under the following conditions (Bridge 1977): (1) there are widespread differences in tastes and these differences are recognized as legitimate; (2) individuals are well informed about quality and different qualities are easily compared and evaluated; (3) purchases are recurrent and relatively small in relation to income so that buyers can learn from experience and easily switch from one brand and supplier to another.

These conditions are ideally present in the case of foodstuffs, but much less so in the case of, say, health and educational services. Hence the development of voice constitutes here an important alternative strategy for assuring and maintaining product quality. In other words, the beneficiaries of certain public services should be induced to become active on their own behalf, individually or collectively As always, development of voice is

arduous because of apathy and passivity of the members, but also because it will often be resisted by the organizations that have been set up to deliver the services. A number of proposals and attempts have been made to introduce more voice into the administration of both health and educational services (Stevens 1974; Klein 1980).

EVL had insisted on the see-saw character of exit and voice interventions in these fields. Education and health systems seemed particularly exposed to the danger that premature exit – of the potentially most influential members – would undermine voice. The opposite relation may also occur, however, for the opening up of the exit perspective could serve to strengthen voice: parents who have been wholly passive because of feelings of powerlessness and fear of reprisals may feel empowered for the first time once they are given vouchers that could be used 'against' the schools currently attended by their children, and will be more ready than before to speak out with regard to desirable changes in those very schools.

Spatial Mobility (Migration) and Political Action

Another substantial area of exit–voice applications opens up when exit is taken in the literal, spatial sense. Here exit–voice boils down to the familiar flight or fight alternative. While often institutionalized among nomadic groups (Hirschman 1981, ch. 11), this alternative is not necessarily available in sedentary societies. Here the traditionally available choice is fight or submit in silence. The option of removing oneself from an oppressive environment has become available on a massive scale only in modern times, with the advances in transportation and the *uneven* opening up of economic opportunity, religious tolerance, and political freedom. Where the option has existed, the interaction of exit and voice has been on display in three principal types of migration: (1) that from the countryside to the city, the oldest and no doubt largest of the modern migrations; (2) the migration from the city to the suburbs, which was most intense in the United States during the fifties and sixties, owing to the spread of the automobile and also to the large-scale migration of blacks and Hispanics *into* the cities; (3) finally, of course, international migration with its numerous economic and political determinants and constraints. Under this rubric, the international movement of capital also deserves attention.

Looking at the varieties of exit-voice interplay in these diverse settings, it is possible, on the basis of the numerous studies now available, to distinguish the following patterns:

(1) In accordance with the basic hypothesis of *EVL,* exit-migration deprives the geographical unit which is left behind (countryside, city, nation) of many of the more activist residents, including potential leaders, reformers, or revolutionaries. Exit weakens voice and reduces the prospects for advance, reform, or revolution in the area that is being left.

Something of this pattern can be observed in all three types of migration. Massive rural-urban migration could obviously reduce the potential as well as the need for land reforms which the voice of the countryside might otherwise have precipitated (Huntington and Nelson 1976, pp. 103ff.). The large outward migration from Europe to the United States in the 19th century up to World War I probably functioned as a political safety-valve for the rapidly industrializing European societies of that period, as has been shown for Italy (MacDonald, 1963–1964). In a similar vein, the possibility of westward migration within the United States has been invoked as an explanation for the lack of a militant working-class movement in that country. Finally, the city-to-suburbs migration in the United States has led, at least initially, to cumulative deterioration in the urban areas affected by out-migration in spite of, and in some cases because of, reduced density. At times, the voice-weakening effect of exit is consciously utilized by the authorities: permitting, favouring, or even ordering the exit of enemies or dissidents has long been one – comparatively civilized – means for autocratic rulers to rid themselves of their critics, a practice revived on a large scale by Castro's Cuba and, on a more selective basis, by the Soviet Union.

(2) But the basic see-saw pattern – the more exit the less voice – does not exhaust the rich historical material. The mechanism through which voice is strengthened rather than weakened as a result of exit is distinctive in the case of migration. In some societies the accumulated social pressures could be so high that authoritarian political controls will only be relaxed if a certain amount of out-migration takes place concurrently. This is what happened in the fifty years prior to World War I when the franchise was extended in many European states from which large contingents of people were departing. In other words, the state accommodated some of the pressures toward democratization because it could be reasonably surmised, in part as a result of out-migration, that opening the door slightly to voice would not blow away the whole structure. A similar positive relation between exit and voice may exist today with regard to such southern European countries as Spain, Portugal, and Greece: here the large-scale emigration to northern Europe may also have eased the transition to a more democratic (more vociferous) order.

(3) Exit–voice theory posits remedial or preventive responses to any large-scale out-migration on the part of the entity that is being left. A firm losing customers or a party losing members will normally undertake a search for the reasons of such declines in fortune and then determine upon a strategy for recovery. For out-migration such reactions are not easy to identify. In the case of massive rural–urban migration, for example, there is usually no organized entity such as the 'countryside' that registers the flight from it and can undertake corrective action. With regard to migration from the city to the suburbs, the situation is not too different. Here entities exist – city administrations – but they have generally been ineffective in modifying the individual decisions of millions of people to move into their own homes in the suburbs.

The analogy to the firm is – or should be – most applicable when the geographic entity losing residents is the State, which is after all a highly organized, self-reflective body with considerable means of action. There is, of course, the already noted possibility that out-migration relieves economic or political stress in a country, is therefore *welcome,* and may even be encouraged by the state. But massive emigration is at some point bound to be viewed as dangerous. Just like a business firm, the state may then take measures to make itself more attractive to its citizens. One example of this reaction is the national plan for economic recovery and industrialization adopted by Ireland in 1958, in the midst of very high levels of emigration, mostly to England (Burnett 1976). It has also been shown that the pioneering welfare state measures of the late 19th and early 20th century, starting in Bismarck's Germany in the eighties and then spreading to the Scandinavian countries and Great Britain, were all taken in countries with high rates of overseas migration. These measures can be seen as attempts of states to make themselves more attractive to their citizens (Kuhnle 1981).

The international movement of capital was first commented upon from the exit perspective in the 18th century. Montesquieu and Adam Smith both thought that the threat of exit on the part of movable capital could play a useful role in preventing arbitrary and confiscatory measures against the legitimate interests of commerce and industry. The threat of exit or exit itself was expected to function, like the customer's exit, as a curb on misconduct, this time on the part of the state. While this relationship is still pertinent, exit of capital often plays a less constructive role today. In the more peripheral capitalist countries the owners of capital have become fully alive to the possibility of removing part of their holdings to the United States or other reliable places in case they become unhappy about the 'investment climate'. In this manner, capital exit (or flight) will often be practised on a large scale as soon as the state undertakes some, perhaps long overdue, reforms with respect to such matters as land tenure or fiscal equity. Instead of preventing arbitrary and ill-considered policies, exit can thus complicate and render more hazardous certain *needed* reforms. Moreover, exit undercuts voice: as long as the capitalists are able to remove their

patrimony to a safe place, they will have that much less incentive to raise their voice for the purpose of making a responsible contribution to national problem-solving. Capital mobility and propensity to exit may thus be a major reason for the instability of states in the capitalist periphery (Hirschman 1981, ch. 11).

Political Parties

Two principal propositions were put forward by *EVL* with regard to the dynamics of political parties in a democracy:

(1) In a two-party system, the tendency of the parties to move toward the nonideological centre in order to capture the (allegedly) voluminous middle-of-the-road vote is countered by those party members and militants who are on the parties' ideological fringes, have 'nowhere else to go', but just because of that are maximally motivated to exert influence inside the party, by forceful uses of voice.
(2) In a multi-party system, with the ideological distance from one part to the next being presumably shorter than in two-party systems, dissatisfaction with party performance is more likely to lead to exit than in two-party systems; in the latter, voice will play the more important role as switching to the other party requires too big an ideological jump. One inference is that parties in two-party systems may be expected to exhibit more internal divisions, but also more internal democracy and less bureaucratic centralism than parties in multiple-party systems.

The first of these propositions has been strongly supported by events subsequent to the publication of the book. At that time, only the nomination of Barry Goldwater to be the standard bearer of the Republican Party in 1964 could be cited in support. Since then, additional evidence has accumulated: from the nomination by the Democrats of George McGovern to contend the Presidential elections in 1972 to the increasing power of the more radical wing of the Labour party and the ascendancy of Margaret Thatcher within the Conservative party and of Ronald. Reagan among the Republicans. The theory that in a two-party system the two parties would increasingly converge toward some middle ground has been amply disconfirmed.

The second proposition on political parties which was deduced from the *EVL* framework has undergone several qualifications. For example, in democracies with old cleavages along ethnic, linguistic, and religious lines, the distances between the several parties rooted in ethnic, etc., identities could actually be wider than that between the parties of two-party systems. Under these conditions, the exit voice logic would in fact predict that member participation (voice) in parties of multi-party systems would also be vigorous and exit infrequent (Lorwin 1971; Hirschman 1981, ch. 9).

A more serious complication is being stressed in a work by S. Kernell still in progress. In two-party systems, exit is a particularly powerful move for dissatisfied members as by casting their vote for the other party they are doubling its impact, something they cannot be sure of in multi-party systems. Hence, in case of disappointment with the performance of one's own party, there could arise a special temptation in two-party systems to switch to the other party so as to *punish* one's own. Such a preference for exit is likely to come to the fore primarily when a party in power is perceived as having seriously mishandled its mandate. Under the circumstances, the prospect of being able to punish that party retrospectively could overcome party loyalty and past ideological commitment. This constellation was an important factor in the sharp defeat of the Democratic ticket in the 1980 Presidential elections in the United States.

The Family: Marriage and Divorce

Modern marriage is one of the simplest illustrations of the exit–voice alternative. When a marriage is in difficulty, the partners can either make an attempt, usually through a great deal of voicing, to reconstruct their relationship or they can divorce. The complexities of the interplay between exit and voice are well in evidence here. Just as the threat of strike in labour-management relations, so is the threat of divorce important in inducing the parties to 'bargain seriously'; but as exit becomes ever easier and less costly (and

perhaps even profitable to one of the parties – see Weitzman 1985), its availability will undermine voice: rather than being an action of last resort, divorce could become the automatic response to marital difficulty with less and less effort made at communication and reconciliation.

This is exactly what appears to have happened in the United States during the last fifteen years, i.e. since *EVL* stated that 'the expenditure of time, money and nerves' necessitated by complicated divorce procedures serves the useful, if unintended purpose of 'stimulating voice in deteriorating, yet recuperable organizations which would be prematurely destroyed through free exit' (p. 79). In 1970 California adopted a new 'no-fault' law on divorce which spread, though often in attenuated form, to most other states (Weitzman 1985). The California law drastically altered divorce procedures: instead of requiring proof that one of the parties was guilty of some specific type of behaviour constituting grounds for divorce, the new law permitted divorce when both *or just one* of the two parties asserted that the marriage had irretrievably broken down. The possibility of a unilateral decision, of just 'walking out', is symbolic of the way in which the California law undercuts the recourse to voice.

With the new regime, the pendulum has swung quite far in the direction of facilitating exit and of thereby weakening voice. It was of course a reaction to the many abuses of the older fault-based system which required costly and degrading adversarial proceedings, and in effect discriminated against the poor. But the framers of the new legislation probably did not realize the extent to which the earlier obstacles to divorce indirectly encouraged attempts at mending the so easily frayed conjugal relationship and how much the new freedom to exit would torpedo such attempts, with the results that one of every two new marriages now ends in divorce.

The Family: Adolescent Development

This is another family situation for whose analysis a formulation in terms of exit and voice has been found useful (Gilligan 1986). Adolescent development has often been portrayed as a process through which the 'dependent' child becomes an 'independent' adult through progressive 'detachment' from the parents. Freud saw this as 'one of the most significant, but also one of the most painful psychic accomplishments of the pubertal period . . . a process that alone makes possible the opposition, which is so important for the progress of civilization, between the new generation and the old' (1905, p. 227). Here is a celebration of exit; Freud's statement neglects a complementary aspect and task of adolescent development which is to maintain and enrich the bond with the older generation through continued, if conflict-ridden, communication. In other words, voice has an important role to play in transforming the adolescent's relationship to the parents. The peculiar poignancy of the adolescent–parents conflict resides in fact in the impossibility of relying *wholly* on voice in resolving it: given the closeness of the relationship, a full accord that would be the outcome of successful voicing risks ending up in incest, as the 'meeting of minds would suggest a meeting of bodies' (Gilligan 1986). It is because of the incest taboo that exit must be part of the solution, but different generations of adolescents are likely to achieve emancipation by practising very different characteristic mixes of exit and voice. Moreover, as Gilligan stresses, the balance of exit and voice differs according to gender. Girls place a greater value than boys on continued attachment to the family, and are therefore less attracted to the masculine ideal of independence-isolation. Hence they experience a greater tension between exit and voice.

With this imaginative use of the exit–voice concept, the outer limits of its sphere of influence may have been reached.

See Also

► Tiebout Hypothesis

Bibliography

Barry, B. 1974. Review article: 'Exit, voice, and loyalty'. *British Journal of Political Science* 4: 79–107.

Bender, P. 1981. *Das Ende des ideologischen Zeitalters*. Berlin: Severin und Siedler.

Birch, A.H. 1975. Economic models in political science: The issue of 'exit, voice, and loyalty'. *British Journal of Political Science* 5: 69–82.

Breneman, D.W. 1983. Where would tuition tax credit take us? Should we agree to go? In *Public dollars for private schools*, ed. T. James and H.M. Levin. Philadelphia: Temple University Press.

Bridge, G. 1977. Citizen choice in public services: Voucher systems. In *Alternatives for delivering public services*, ed. E.S. Savas. Boulder: Westview.

Burnett, N.R. 1976. *Emigration and modern Ireland*. Unpublished PhD dissertation, School of Advanced International Studies, Johns Hopkins University.

Eccles, R.G. 1981. The quasifirm in the construction industry. *Journal of Economic Behavior and Organization* 2 (4): 335–357.

Fainstein, N.I., and S.S. Fainstein. 1980. Mobility, community, and participation: The American way out. In *Residential mobility and public policy*, ed. E.G. Moore and W.A.V. Clark. Beverly Hills: Sage.

Finer, S.E. 1974. State-building, state boundaries and border control in the light of the Rokkan-Hirschman model. *Social Science Information* 13 (4–5): 79–126.

Freeman, R.B., and J.L. Medoff. 1984. *What do unions do?* New York: Basic Books.

Freud, S. 1905. Three essays on the theory of sexuality. In *Complete psychological works*, vol. 7. London: Hogarth, 1953.

Gilligan, C. 1986. Exit–voice dilemmas in adolescent development. In *Development, democracy and the art of trespassing: Essays in honor of A.O. Hirschman*, ed. A. Foxley et al. Notre Dame: University of Notre Dame Press.

Granovetter, M. 1985. Economic action and social structure: A theory of embeddedness. *American Journal of Sociology* 91: 481–510.

Hirschman, A.O. 1970. *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Cambridge, MA: Harvard University Press.

Hirschman, A.O. 1977. *The passions and the interests: Political arguments for capitalism before its Triumph*. Princeton: Princeton University Press.

Hirschman, A.O. 1981. *Essays in trespassing: Economics to politics and beyond*. Cambridge: Cambridge University Press.

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 some categories of economic discourse. *Economics and Philosophy* 1: 7–21.

Huntington, S.P., and J.M. Nelson. 1976. *No easy choice: Political participation in developing countries*. Cambridge, MA: Harvard University Press.

Kernell, S. 1987. *Retrospective voting and contemporary macrodemocracy*. Washington, DC: Brookings Institution.

Klein, R. 1980. Models of man and models of policy: Reflections on exit, voice and loyalty ten years later. *The Milbank Memorial Fund Quarterly* 58: 413–429.

Kuhnle, S. 1981. Emigration, democratization, and the rise of the European welfare states. *Mobilization, center-periphery structures, and nation-building,* (a volume in commemoration of Stein Rokkan), ed. P. Torsvik. Bergen: Universitetsforlaget.

Levin, H.M. 1983. Educational choice and the pains of democracy. In *Public dollars for private schools: The case for tuition tax credits*, ed. T. James and H.M. Levin. Philadelphia: Temple University Press.

Lorwin, V. 1971. Segmented pluralism: Ideological cleavages and political cohesion in the smaller European democracies. *Comparative Politics* 3 (2): 141–175.

MacDonald, J.S. 1963–1964. Agricultural organization, migration and labour militancy in rural Italy. *Economic History Review* 16: 61–75.

O'Donnell, G. 1986. On the convergences of Hirschman's exit, voice and loyalty and shifting involvements. In *Development, democracy and the art of trespassing: Essays in honor of A.O. Hirschman*, ed. A. Foxley et al. Notre Dame: University of Notre Dame Press.

Rokkan, S. 1975. Dimensions of state formation and nation-building: A possible paradigm for research on variations in Europe. In *The formation of national states in Western Europe*, ed. C. Tilly. Princeton: Princeton University Press.

Stevens, C.M. 1974. Voice in medical care markets: 'consumer participation'. *Social Science Information* 13 (3): 33–48.

Weitzman, L.J. 1985. *The divorce revolution: The unexpected social and economic consequences for women and children in America*. New York: Free Press.

Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Expectations

Robert J. Shiller

Abstract

The modelling of economic expectations is central to economics. Expectations of future economic conditions can be represented in econometric models by survey data, expectations proxies such as adaptive expectations,

expert forecasts, or market expectations. The theory that expectations are rational, that is, optimal forecasts given the model, can be a useful modelling device, but evidence from behavioural economics shows that it has important limitations.

Keywords

Adaptive expectations; Aspirations; Behavioural economics; Capital asset pricing model; Certainty equivalent; Distributed lag; Error-learning hypothesis; Euler equations; Expectations; Expectations proxies; Expected utility; Extrapolative expectations; Forecasting; Iowa Electronic Market; Law of iterated projections; Market expectations; Mathematical expectations; Muth, John F.; Phillips curve; Phillips, A. W.; Prediction markets; Quadratic expected utility function; Rational expectations; Rational expectations equilibrium; Regressive expectations; Stochastic optimal control; Subjective probability; Survey expectations; Transformations of variables; Unexpected inflation; Variance

JEL Classifications

D84

Most decisions that economic agents must make involve uncertainty about the future. Thus, any economic model that is intended to be descriptive of human behaviour is likely to involve human expectations about uncertain future economic variables. Areas in economics that involve expectations in fundamental ways include the theories of intertemporal consumption or labour supply decisions, theories of firms' pricing, sales, investment, or inventory decisions, theories of financial markets and money, theories of insurance, and of search behaviour, signalling, agency and bidding. If our purpose is to describe human behaviour, then the study of human expectations is inseparable from the study of the behavioural models in which these expectations are embedded. Only a few general observations can be discussed here.

Economic Expectations, Surveys and Proxies

Applied econometric research often relies on simple models involving *expectations variables* that represent the expectations of economic agents for some specified economic variables. For example, the total savings of individuals may be related to a variable purporting to measure their expectations for their pension benefits on the date of their retirement, years in the future. Or an individual's decision whether to purchase a long-term or short-term bond may be related to a variable representing expectations as to the course of future short-term interest rates over the life of the long-term bond.

Expectations variables included in such models are often referred to as measuring, perhaps imperfectly, some idealized *economic expectations.* What economic expectations actually represent is usually not spelled out. Different people have different perceptions about the outlook for future variables, and so there is an index number problem in reducing their divergent opinions into a single measure. Moreover, when asked for their expectation for some economic variable people may answer that they have no expectation. If pressed, they may hazard a guess. Certainly, most individuals make some economic decisions without making an effort to learn about relevant economic variables. From time to time, circumstances require making difficult or important decisions, and then people may trouble themselves more to find out about economic variables. Economic models that speak of 'the' expectation of an economic variable presumably are talking about some average of the expectations of some people and guesses other people would make if pressed, or about averages of the better or worse forecasts of the same people at different times.

The expectations variables used in econometric work to measure economic expectations may be *survey expectations,* representing the average expectation respondents reported on a public opinion survey, or they may be *expectations proxies*, consisting of transformations of other variables that appear to the econometrician to be

plausible guesses as to the public expectations. For example, a moving average of lagged inflation rates may serve as an expectations proxy for future inflation.

Expectations surveys commonly take two forms: those that survey individuals representative of the general population and those that sample experts. The former provide measures of expectations that are relevant to decisions, like individual decisions as to how much to save in a given month or whether to put money in a savings account as against corporate bonds, on which decision-makers do not attach great importance. The latter provide measures of expectations that are relevant to decisions, such as firm decisions on whether to market a new product or invest in a new plant, on which decision-makers are likely to spend the resources to obtain informed forecasts.

Day-To-Day Expectations of the General Population

According to Katona (1975), survey research finds that most people can be induced to make a guess as to the direction of change in the near future of major macroeconomic variables, but are reluctant to give quantitative estimates of the extent of the change. The information on which most people base their expectations is fragmentary. Based on decades of survey research on the general public in the United States, Katona concluded that the majority knew whether unemployment had increased or decreased in the preceding months, whether profits or retail sales had gone up or down, and also whether interest rates had risen or fallen, but did not know how much larger or smaller any of these magnitudes were. The extent of knowledge about macroeconomic variables is generally greater the more important or dramatic the recent changes in these variables, and of course, the more the variable has been emphasized in the mass media.

Since we generally want to incorporate expectations variables in an economic model that describes human behaviour, we are likely to want any variable measuring economic expectations to represent the actual thoughts of individuals *before* they were forced to sit down at a questionnaire and carefully think about how to forecast an economic variable. In modelling, say, income expectations for the purpose of studying the saving decision, we want to get into the individual's frame of mind at the times when saving decisions are made.

When we try to characterize a person's frame of mind at these times, we should recognize that the expectations are likely to differ through time qualitatively as well as quantitatively. For example, an expectation of a future rise in income may become more vividly impressed on individuals' consciousness by some public event that reminds that person of the reasons to expect income to rise. At the same time, the expectation as measured on a survey may be unchanged. Psychologists who study the saving decision have emphasized the importance of changing *aspirations* as distinct from changing expectations.

Individuals who are not thinking at all about economic theories and who are merely confronted with economic variables whose stochastic properties are difficult to comprehend may fall back on simple expectations mechanisms such as that proposed by Fisher (1930). In his, expected inflation is a *distributed lag* or weighted average, with weights that decline linearly with time, of actual inflation. A variation on Fisher's expectation mechanism is *adaptive expectations* (Cagan 1956) in which expectations are formed as a weighted average, with weights that decline exponentially with time into the past, and that sum to 1, of actual past inflation. The rate at which weights decline might be determined by the rate at which human memory decays. It may be natural to form expectations of a future variable (for example, inflation) by thinking back over the recent past of experience of the variable, and hence such memory decay may result in a distributed lag pattern like that hypothesized by Cagan.

With adaptive expectations, the change in the expectations variable is proportional to the difference between its previous value and the latest value of the variable to be forecasted. This construction resembles that of the *error-learning hypothesis* (Meiselman 1962). However, in the error learning hypothesis the change in the

expectation for a variable at a specific future date is proportional to the error just discovered in the forecast for the variable for today's date.

Alternatives to adaptive expectations are *regressive expectations*, in which variables are expected to return gradually to a fixed level independent of their recent past behaviour (this term has also been used as a synonym for adaptive expectations), and *extrapolative expectations* in which the recent direction of change in the variables is expected to continue (see for example Modigliani and Sutch 1966.) Any of these expectations mechanisms may be consistent with optimal forecasts of the future variables under certain special circumstances (see Sargent and Wallace 1973).

We may not want to use such simple models of expectations in periods when individuals may think a great deal about economic theories. During a period of hyperinflation, for example, it is perhaps unlikely that people will form expectations adaptively, since the inflation affects them so noticeably. They may seek out the opinions of experts at such times.

Expectations of Experts

It is often the case that it is much easier for surveyors to find the expectations of randomly sampled individuals than the expert opinions. Expert opinions may be generated only at the time a crucial decision is made, and not when an expert is asked to fill out a questionnaire. Moreover, experts may feel that their time is too valuable to merit attending carefully to a questionnaire.

It is now the case that economic forecasting has become a profession in which practitioners regularly publish their forecasts of macroeconomic variables, and thereby open themselves up to systematic evaluation by outsiders. Usually these forecasts have some basis in econometric models subject to judgmentally introduced 'add factors'.

Professional forecasters now make available regularly tabulated forecasts of macroeconomic variables for the succeeding few years. The accuracy of these forecasts is now regularly computed by independent evaluators, and this provides a genuine incentive to forecast well. The marketplace will tend to reduce the numbers of those who do not forecast well. Professional forecasts made in organizations, when not made in anticipation of the kind of 'forecasting race' judged by outside evaluators, may not be serious individual attempts to predict. Instead, they may be 'conventional' forecasts using methods and information that are perceived as having sanction in the organization. Organizations may stipulate what information a forecaster is to use and how the information is to be translated into a forecast. The aim of such sanctions may be to produce uniformity in the organization as to factual premises on which decisions are made, but they may also lead to forecasts that are not accurate. The costs to individuals of violating the assumptions of the organization may be very large relative to the possible benefits of forecasting well.

The distinction between day-to-day expectations of individuals and the expectations of experts may in practice not be an important one. The advantages that experts have, of access to data, understanding of economic theory and use of statistical methods, may confer little advantage in circumstances when the structure of the economy is changing. Then the data may be viewed as of little help, as it is generated by a different model, and statistical analysis also may be of little help. Experts may then fall back on adaptive expectations or other methods of producing guesses like those of ordinary individuals.

Mathematical Expectations

When we use the term *mathematical expectations,* we are referring to a probabilistic model, from which we can compute the expectations as first moment conditioned on the information set available to agents. There are of course other candidates to represent economic expectations, for example, other measures of central tendency such as the median or mode, or measures of central tendency applied to transformations of the random variable.

Ultimately, many of economic models that involve mathematical expectations as economic

expectations derive from the assumption of maximization of the mathematical expectation of a utility function. The mathematical expectations operator is initially brought into the assumptions of the model because such expected utility maximization is viewed as a good way to represent human behaviour. Expected utility maximization has been shown to follow from some plausible axioms representing an idealization of 'rational' human behaviour. But it is only in certain special cases that maximization of expected utility produces simple behavioural relations involving mathematical expectations as 'economic expectations' of the kind that many applied econometricians have been using.

Linear utility functions representing risk-neutral agents may give rise to models in which agents care only about the mathematical expectations of variables, as in the models in finance in which the mathematical expectations of returns on various assets are equalized. A quadratic expected utility function may also produce models that depend on mathematical expectations. It is a result of Simon (1956) that if there are no terms of degree higher than two in control variables and exogenous stochastic processes then optimal behaviour depends linearly on a 'certainty equivalent' equal to the conditional expected value of future values of the stochastic process, and not on any other characteristics of their conditional distribution. Simon set up a problem in which there was nothing that could be done by the maximizing agent about the variance of the outcome. In contrast, in the capital asset pricing model in finance, a utility function quadratic in wealth (but where there are terms of degree higher than two in control variables and wealth) yields a behavioural relation that involves both a mathematical expectation and a variance matrix of the underlying stochastic variables.

More generally, expected utility function models that are not linear or quadratic will produce Euler-equation type first-order conditions involving the mathematical expectation operator and economic variables. These models can then give rise to relations involving mathematical expectations that may be interpreted as economic expectations. Suppose economic agents at time t maximize subject to a budget constraint an intertemporal utility function $\sum_{k=0}^{\infty} \beta^k u(C_{t+k})$ where β is the subjective discount factor, u represents instantaneous utility and C_{t+k} represents consumption k periods after time t. Then the Euler equation implied by agents' optimization states that, for any liquid asset whose price is P_t at time t that pays no dividend between t and $t+1$, $P_t = E_t\left(\beta \frac{u'(C_{t+1})}{u'(C_t)} P_{t+1}\right)$ where E_t denotes mathematical expectation. If we can find some reason to assume that the term $\beta \frac{u'(C_{t+1})}{u'(C_t)}$ can be disregarded, such as by assuming either that the time interval is small and that there is little variation in consumption over this time interval, then the price P_t itself can be interpreted as the mathematical expectation of price P_{t+1} next period.

Many models start from behavioural relations involving mathematical expectations and do not derive these from the hypothesis of expected utility maximization. In these cases, the popularity of mathematical expectations as representations of economic expectations may derive from some intuitively desirable and convenient properties of mathematical expectations, properties that are not shared by other measures of central tendency. The mathematical expectation of the sum of two random variables is equal to the sum of their mathematical expectations whether or not the two variables are independent, a property not shared by the median or mode, even if the variables are independent. If we have a joint distribution of two random variables, x and y, and we define the conditional distribution of x given y, then the mathematical expectation $E(x|y)$ of x in the conditional distribution is a function of y. The *law of iterated projections* states that the mathematical expectation of the mathematical expectation of x, $E(E(x|y))$ equals the mathematical expectation of x, $E(x)$. In simple terms, this law might be described as saying that people do not expect to change their expectations. Again, this law does hold in general for the mode or median of x. On the other hand, the median has the desirable property that the median of any monotonic transformation of a random variable is the transformation of the median, a property not shared by the mathematical expectation.

Market Expectations

In 1907, the statistician Francis Galton did a statistical analysis of a contest in which participants paid, for the possibility of a monetary prize, to play a game in which they guessed the weight of an ox. His conclusion that the average guess was very close to the actual weight of the ox led over the years to a general public appreciation of the idea that markets may predict very well, and hence that market expectations represent optimal forecasts.

Economic theory may, under some conditions, support the idea that financial market prices may represent mathematical expectations conditioned on public information. If we have a security whose value in the near future P_{t+1} unambiguously represents some economic value that is highly variable, then we might conclude that its price today is its mathematical expectation. (If we assume itis variable relative to the potential variability in $\beta \frac{u'(C_{t+1})}{u'(C_t)}$in the Euler equation, then the price today P_t might be regarded as approximately the mathematical expectation of that economic value.) On the other hand, if the security represents a claim on the distant future, then these assumptions seem problematic, and even the validity of the Euler equation itself becomes questionable (Shiller 2005).

The Iowa Political Stock Market was created in 1988 by Robert Forsythe, Forrest Nelson and George Neumann at the University of Iowa to allow participants to buy securities that pay one dollar plus the candidate's final vote margin in an election in the near future. The price of such a security may be interpreted under our assumptions as 1 plus the mathematical expectation of the candidate's vote margin.

Their market, now renamed the Iowa Electronic Market, also trades securities that pay one dollar if an event occurs, otherwise nothing. Then the price of that security has a possible interpretation as the probability that the event will occur.

In fact, of course, not everyone has the same subjective probability of the event. The differences of opinion may be necessary for a market to function, for it may be only the differences of opinion that make the market interesting to traders. Manski (2006) shows that there is no theoretical support for the idea that the market prices should represent the average (over all market participants) subjective probability of the event, and, according to his assumptions, theory allows us to define only a (rather broad) interval within which this average subjective probability lies. Still, the market prices might turn out to be useful in helping us to judge probabilities of future events (Wolfers and Zitzewitz 2006).

The Iowa people have claimed that the prices of their securities representing final vote margins have produced remarkably accurate forecasts, better than that generated by public opinion polls of voter intentions (Berg et al. 2008). However, the economics profession has not yet generated many studies that test the interpretation of these new markets as generators of optimal expectations.

Interest in market expectations remains very high, and there has been an explosion of prediction markets. Some are related to universities: the Austrian Electronic Market run by Vienna University; the University of British Columbia Election Stock Market. Others are private companies such as intrade.com; cityindex.co.uk; igindex.co.uk; tradesports.com; hedgestreet.com; newsfutures.com; and ideosphere.com.

The Economic Derivatives Market was created in 2002 by Deutsche Bank and Goldman Sachs to trade claims that pay out a fixed sum if an economic variable falls in a specified range, using a trading platform created by Longitude, Inc. that allows people to express complex demands for this security (Lange and Economides 2005). The market is now managed in a partnership between Goldman Sachs and the Chicago Mercantile Exchange. Today US non-farm payrolls, the Institute of Supply Management's Purchasing Managers Index (PMI), weekly initial jobless claims, retail sales, the European harmonized index of consumer prices (HICP), the international trade balance and gross domestic product (GDP) are currently traded. Preliminary studies, with only limited data available as yet, support the notion that these markets yield useful forecasts (Gurkaynak and Wolfers 2006).

In 2006 the Chicago Mercantile Exchange with MacroMarkets LLC (a company I helped found) created futures and options markets that are cash-settled based on the Standard & Poor's/ Case-Shiller Home Price Indices, and has plans to create such markets based on a commercial property price index. These markets are beginning to offer market expectations for real estate prices.

Other examples of potential new markets for economic variables are described in Shiller (2003). The value of these markets for generating market expectations or optimal forecasts will not be in until many more years' data are at hand.

Modelling Rational Expectations

Even if we can use markets or surveys to measure expectations, if we are to understand their movements through time we need to model their determination. The idea that expectations are rational has been an important modelling device.

Why does the idea sound plausible that economic expectations of future inflation may be proxied fairly well by adaptive expectations or other distributed lag on actual inflation? Is it just because of the theory of psychologists that human memory decays gradually through time and the notion that casual guesses of future inflation would correspond to recent memories of inflation? Perhaps it is instead that a distributed lag on inflation is not a bad way to forecast inflation.

Suppose people were asked on a monthly basis to forecast the rate of increase of the price of some seasonal commodity, let us say, fresh tomatoes. Certainly, many of them would be aware that fresh tomatoes are more expensive in the winter, when they must be grown in hothouses or brought in from greater distances. Not all people would know this, and many who did know about the seasonality in price would not know its magnitude. But certainly a distributed lag with smoothly declining coefficients on actual tomato price changes is not what we would think of first to model their expectations. Such a distributed lag would imply some seasonality in expectations but would also generally imply that people misforecast the month of highest price.

If there is any doubt as to the value of simple expectations proxies for modelling the expectations of tomato consumers, there is certainly no doubt that it would be inappropriate to use such proxies to model the expectations of tomato producers. Some producers specialize in producing hothouse tomatoes, and time their production for the winter months. Surely they know in which month prices are higher, and by how much they tend to be higher.

How then should we build a model that describes the supply of tomatoes over time? Since tomatoes must be planted months in advance of the anticipated demand for them, the supply function for tomatoes must depend on expectations formed at this time by producers, as well as on seasonal factors affecting the cost of production. We might then model the supply of tomatoes by finding a good way to predict the price of tomatoes (using, say, seasonal dummies and other information) and substituting the prediction in place of the expectation in the supply function. The result would be a *rational expectations model.*

One could use such a model to predict the supply response to some variable that has been found to predict price. If, let us say, we found that bad weather in Mexico, which might later reduce supply of winter tomatoes to the United States, tended to cause the seasonal peak in tomato prices in the United States to be higher than usual, then we might in these circumstances forecast the supply of domestic tomatoes in the United States to be higher than usual. A rational expectations model would produce such a forecast if the model was based on an empirical forecasting relation for price that used the weather variable as an explanatory variable.

Of course, for the purpose of forecasting supply we might also have used the 'naive' approach of estimating a forecasting equation directly for supply (without the intermediate step of developing a forecasting equation for price) depending on such variables as earlier weather and on seasonal dummies. Such a method may also satisfactorily predict supply, but it might not do as well since it would not make use of the information in economic theory that weather affects supply only

through its effect on rationally expected price. For example, suppose we had a long time series of data on various weather variables and prices but only a few observations on quantities supplied. We could not include all the weather variables directly in a 'naive' forecasting equation for supply, since we would thereby exhaust degrees of freedom. But we could first find how these weather variables predict price and then use a single price expectations variable to predict supply.

Rational Expectations in Equilibrium Models

The above example of the use of a rational expectations model was very special in that the model consisted only of a single equation relating supply to an earlier expectation of price. Moreover, the equation was used only to forecast supply in a situation where we expect the correlations observed in the past with explanatory variables to continue. Very often we wish instead to predict the effect on supply of some change in government policy or other structural change that is expected to change the correlations with other variables.

Suppose for example we wish to know the effect on the seasonal pattern of tomato supply in the United States of a government policy of blocking the further international trade of tomatoes. Here, the naive forecasting model that related tomato supply to weather and seasonal dummy variables would be of no value. An estimated rational expectations model relating supply to expected price might still be of value. We need to model only the determination of expected price.

Suppose we then also estimate (using a sample period in which some tomatoes were imported) a domestic demand function for tomatoes, relating, say, total quantities demanded in the United States to contemporaneous price. Consider, then, a two-equation model consisting of this demand equation and the rational expectations domestic supply equation for tomatoes described above. In the sample period domestic demand did not equal domestic supply because of imports. After the policy change the domestic supply and demand will be equal. Can we now predict how the seasonal pattern of quantities supplied may be changed by the government policy?

To answer this question, we cannot just solve the two-equation model with the two endogenous variables, quantity and price, because both price and expectations appear separately in the model. However, the expectation of price, if it is a rational expectation, ought to be determined by the very model in which it appears. How can we find the rational expectation of price?

One approach is first to guess a function relating expected price to the exogenous variables in the model, in our example, the seasonal dummies and weather variable. If one substitutes this guess into the model in place of the expected price, one then has an ordinary simultaneous equation model in price and quantity in terms of exogenous variables. However, unless one made a lucky guess, one would then find that the model that resulted from the guess was inconsistent with the guess, in that the model implies that a different way of forecasting price is optimal, given the expectations function.

What we need to find is an equation defining the expectation of price which, on substituting into the model, produces a model in which that equation gives the optimal forecast of price. Muth (1961) showed how this can be done if the simultaneous equations model is linear and if rational expectations are defined as mathematical expectations conditioned on variables in the model that are in the public's information set.

Using such a solution method, we might find how the seasonality of both quantities and price will be changed under the new government policy. In this simple example, doing this would seem to be preferable to using a model with an expectations proxy for price that did not take into account how the changing seasonal pattern of price would change the way expectations are formed.

Rational Expectations Models, Stochastic Processes and Optimal Control

The advent of rational expectations in econometric models has marked a revolution in economic thinking that is comparable in the magnitude of

its impact on the economics profession to the Keynesian revolution in the mid-twentieth century.

Muth (1961) and those who carried on the rational expectations literature have borrowed heavily from another literature that was once outside economics, namely the theory of stochastic processes and optimal control. What is substantially new about the rational expectations models derives ultimately from these theories, which were developed for the most part since 1950. The implications of these theories were so profound that it was inevitable that they should make themselves felt in economics, just as they have in many fields in science and engineering.

The rational expectations revolution is not primarily the result of any failure of conventional econometric models to forecast well, as some (for example, Lucas and Sargent 1981) have argued. It is true that initial optimism for the forecasting ability of such models has been tempered by experience, but it has not been established that shortcomings of the expectations modelling methods has been the major fault. It has certainly not been established empirically that rational expectations models can predict better.

Interest by economists in optimal control and the theory of stochastic processes was initially expressed in their efforts to apply control methods to existing econometric models, to achieve their stabilization. However, the optimal control of conventional 'Keynesian' econometric models involving expectations proxies like adaptive expectations has never became as influential in the profession as its developers had hoped. Perhaps the general profession thought that the methods of control were too refined for the crude models that they were applied to. More concern was felt for improving the models themselves.

The idea that optimal control might be applied to conventional Keynesian econometric models did have the effect of generating hopes that the macroeconomy might be controlled very well, 'fine tuned' so to speak, and thus great importance was placed on the structural stability of these models. Much of the polemics against 'Keynesian' economics waged by those who promoted rational expectations models as alternatives was really directed against these efforts to apply optimal control systematically to the models (see for example Sargent and Wallace 1981). A central criticism of these models was their heavy reliance on crude expectations proxies such as adaptive expectations.

The rational expectations models applied stochastic optimal control theory by assuming in effect that human behaviour could be modelled as if everyone all along had been applying the principles of optimal control to their own economic decisions. Given the natural interest of economists in rational behaviour, the optimal filtering and extrapolation that was developed as part of the theory of stochastic processes would naturally be used in modelling how individuals forecast.

Of course, there are strict limits to the extent to which people's actual behaviour can be described in such terms. Rational expectations models thus often sacrifice descriptive accuracy with the hope that the models would exhibit stability in the presence of interventions of the kind envisioned by makers of government macroeconomic policy. The models may not be generally well suited to forecasting when the policy regime is unchanged. They are most appropriately considered as policy analysis tools.

Criticisms of Rational Expectations Models

The simple supply and demand model for tomatoes described above was chosen as an ideal example of the application of rational expectations models. In this example there is substantial seasonal variation in price, which ought to be forecastable. Moreover, as the model was set up, only producers' expectations entered the model, and producers are far more likely than others to have rational expectations about price. But few of the applications of the theory of rational expectations have been to such ideal examples.

The best-known application of rational expectations models has been to an interpretation of the observed relation between unemployment and inflation. A.W. Phillips (1958) noted a negative

relation between the unemployment rate and the wage inflation rate in the United Kingdom between 1861 and 1957. A similar relation was found also in the United States for much of the same sample period. Since then, the negative relation has broken down. Lucas (1976) and Sargent and Wallace (1973) offered interpretations of the Phillips relation and its subsequent breakdown. In its simplest terms, this interpretation asserts that there may be a stable relation between unemployment and *unexpected* inflation. Unexpected inflation may cause job seekers to misperceive the real value of wage offers they have received, and thus to accept offers that they would not have accepted if they had known the true real wage they were getting. By accepting these jobs, they lower the unemployment rate. In the period Phillips studied the price level might have been well-enough approximated by a random walk that actual inflation may have approximately equalled unexpected inflation. Since then, when inflation has become much more serially correlated, actual and expected inflation may have diverged widely.

In its general idea, the Lucas–Sargent–Wallace theory of the Phillips curve sounds like an appealing possibility. The question for econometric testing of the theory is whether we want to assume that expectations of unemployed workers are fully rational.

The tests Sargent (1976) made of the model are illustrative of the manner in which rational expectations models are often tested. Sargent tested whether the model holds under the assumption that unemployed workers are making optimal use in their forecasts of inflation of current and lagged values of the real government surplus, real and money government expenditures, the price level, the money supply, and a wage index. It is commonplace today in the rational expectations literature to see similar extravagant assumptions about the information sets of ordinary individuals.

The most basic criticism of many rational expectations models is that they make implausible claims for individual economic agents' ability and willingness to compute. But the criticism of these models goes beyond that: see for example Friedman (1979), Tobin (1980).

The rational expectations models assume that economic agents behave as if they know the structure of the economy so that they can compute the optimal forecasts that represent their expectations. But the structure of the economy is always changing, as technology, tastes and government interventions change. These changes themselves vary qualitatively from time to time, and so it may not be possible for economic agents to group instances in such a way as to allow dealing with the changes in statistical terms. If these changes occur frequently relative to the speed at which people can figure out the economy, it may never be appropriate to assume that their forecasts are optimal forecasts.

In most rational expectations models, the behaviour of the economic variables that individual economic agents must forecast is itself affected by the way the economic agents form expectations. This fact was noted above in connection with our efforts to solve the supply and demand rational expectations model for tomatoes. Thus, if economic agents learn something about how to forecast an economic variable, the random properties of the economic variable may change in consequence. A rational expectations equilibrium is achieved only when people have adopted a way of forecasting that is consistent with the implications for the economy of their own way of forecasting. How do they find such a way of forecasting? Achievement of a rational expectations equilibrium might take place as a consequence of a long iterative process, each step representing the learning by economic agents of how to forecast in the preceding step, and thereby necessitating the next step of learning anew how to forecast. In models that are more complicated than the simple supply and demand models, for example, models of the entire macroeconomy, the time required for each step may need to be enormous. The problem of convergence of forecasting methods to a rational expectations equilibrium recalls the problem in mathematical economics of the convergence of a price vector to Walrasian equilibrium. However, the former problem has received much less attention. Moreover, convergence may well be orders of magnitude slower in the former. It would appear likely, given the

complexity of the macroeconomy, that economic agents learn very slowly about how to forecast given the present structure of the economy. Each step in the iteration requires sifting through large amounts of data and learning how these are related statistically.

The behavioural economics revolution, still very much under way since its beginnings around 1980, can be described as groping for alternatives to rational expectations models, alternatives that have some structure from research in psychology and that do not impose unrealistic complexity on individual decisionmaking. Akerlof and Yellen (1985) have argued for 'near-rational expectations' models. Richard Thaler (1991) has argued that we must turn to something he calls 'quasi-rational economics'. Frydman and Goldberg (2007) have argued that we must work towards something they call 'imperfect knowledge economics'. These are important beginnings, though today none of the alternatives has yet won widespread acceptance among economists.

Despite the criticisms, rational expectations models may well be useful for some applications when compared with alternative models based on expectations proxies. As regards the assumptions in the models for the ability and willingness of economic agents to store and process information, there is no alternative for model builders to that of judging for plausibility on a case-by-case basis.

See also

- ▶ Adaptive Expectations
- ▶ Behavioural Economics and Game Theory
- ▶ Certainty Equivalence
- ▶ Prediction Markets
- ▶ Rational Expectations

Bibliography

Akerlof, G.A., and J.L. Yellen. 1985. A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100: 823–838.

Berg, J., R. Forsythe, F. Nelson, and T. Reitz. 2008. Results from a dozen years of election futures markets research. In *Handbook of experimental economic results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

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

Friedman, B.M. 1979. Optimal expectations and the extreme information assumptions of 'rational' expectations models. *Journal of Monetary Economics* 5: 23–41.

Frydman, R., and M.D. Goldberg. 2007. *Imperfect knowledge economics: Exchange rates and risk*. Princeton: Princeton University Press.

Galton, F. 1907. Vox populi. *Nature* 75: 450–451.

Gurkaynak, R., and J. Wolfers. 2006. *Macroeconomic derivatives: An initial analysis of market-based macro forecasts, uncertainty, and risk*, Working paper no. 11929. Cambridge, MA: NBER.

Katona, G. 1975. *Psychological economics*. New York: Elsevier.

Lange, J., and N. Economides. 2005. A parimutuel market microstructure for contingent claims. *European Financial Management* 11: 25–49.

Lucas, R.E. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.

Lucas, R.E. Jr., and T.J. Sargent. 1981. After Keynesian macroeconomics. In *Rational expectations and econometric practice*, ed. R.E. Lucas Jr. and T.J. Sargent. Minneapolis: University of Minnesota Press.

Manski, C. 2006. Interpreting the predictions of prediction markets. *Economics Letters* 91: 425–429.

Meiselman, D. 1962. *The term structure of interest rates*. Englewood Cliffs: Prentice Hall.

Modigliani, F., and R. Sutch. 1966. Innovations in interest rate policy. *American Economic Review, Papers and Proceedings* 56: 178–197.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25: 283–299.

Sargent, T.J. 1976. A classical macroeconomic model for the United States. *Journal of Political Economy* 84: 207–237.

Sargent, T.J., and N. Wallace. 1973. Rational expectations and the dynamics of hyperinflation. *International Economic Review* 14: 328–350.

Sargent, T.J., and N. Wallace. 1981. 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. In *Rational expectations and econometric practice*, ed. R.E. Lucas and T.J. Sargent. Minneapolis: University of Minnesota Press.

Shiller, R.J. 2003. *The new financial order: Risk in the 21st century*. Princeton: Princeton University Press.

Shiller, R.J. 2005. *Irrational exuberance*. 2nd ed. Princeton: Princeton University Press.

Simon, H.A. 1956. Dynamic programming under uncertainty with a quadratic objective function. *Econometrica* 24: 74–81.
Thaler, R.H. 1991. *Quasi-rational economics*. New York: Russell Sage Foundation.
Tobin, J. 1980. *Asset accumulation and economic activity.* Yijo Jahnsson Lecture. Oxford: Basil Blackwell.
Wolfers, J., and E. Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives* 18 (2): 107–126.
Wolfers, J., and Zitzewitz, E. 2006. *Interpreting prediction markets prices as probabilities*. Working paper, Rodney White Center for Financial Research. Wharton School: University of Pennsylvania.

Expected Utility and Mathematical Expectation

David Schmeidler and Peter Wakker

Expected utility theory deals with choosing among acts where the decision-maker does not know for sure which consequence will result from a chosen act. When faced with several acts, the decision-maker will choose the one with the highest 'expected utility', where the expected utility of an act is the sum of the products of probability and utility over all possible consequences.

The introduction of the concept of expected utility is usually attributed to Daniel Bernoulli (1738). He arrived at this concept as a resolution of the so-called St Petersburg paradox. It involves the following gamble: A 'fair' coin is flipped until the first time heads up. If this is at the *k*th flip, then the gambler receives $\$2^k$. The question arose how much to pay for participation in this gamble. Since the probability that heads will occur for the first time in the *k*th flip is 2^k (assuming independence of the flips), and the gain then is $\$2^k$, the 'expected value' (i.e. the *mathematical expectation* of the gain) of the gamble is infinite. It has been observed though that gamblers were not willing to pay more then \$2 to \$4 to participate in such a gamble. Hence the 'paradox' between the mathematical expectation of the gain, and the observed willingness to pay.

Bernoulli suggested that the gambler's goal is not to maximize his expected gain, but to maximize the expectation of the logarithm of the gain which is $\sum_{j=1}^{\infty} 2^{-j}\log 2^j$, i.e. $2\log 2 (= \log 4)$. Then the gambler is willing to pay \$4 for the gamble. The idea that *homo economicus* considers the expected utility of the gamble, and not the expected value, is a cornerstone of expected theory.

In the next section the approach of Savage to decisions under uncertainty is presented. In section "Expected Utility when Applied to Decisions under Risks; The Von Neumann–Morgenstern Approach" the von Neumann-Morgenstern characterization of expected utility maximization for the context of decisions under risk is given. Section "Other Approaches and Bibliographical Remarks" briefly mentions some related approaches. Section "Appendix: Mathematical Expectation", the Appendix, defines (mathematical) expectation.

Expected Utility when Applied to Decisions Under Uncertainty; Savage's Approach

The main ingredients of a decision problem under uncertainty are acts consequences and states of nature. Suppose that a decision-maker has to choose one of three feasible acts *f, g, h*. Act *f* leads to one (only) of the two consequences a and b. Act *g* leads to a or c, act *h* to b or d. Thus the set of consequences, *C*, is in this example {a, b, c, d}.

The matching of feasible acts to consequences is expressed by the concept of 'state of nature', or 'state' for short. More precisely, a given state of nature indicates for each feasible act what the resulting consequence will be. In the above example, there are three feasible acts *f, g, h*, each leading to one of two possible consequences. See Table 1.

A state of nature completely resolves the uncertainty relating acts to consequences. If the decision-maker would know for sure which state of nature is the true one, then he would choose an act which results in a most desirable consequence.

Expected Utility and Mathematical Expectation, Table 1 The eight logically possible matchings of feasible acts to *consequences*

	States							
Acts	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
f	*a*	*a*	*a*	*a*	*b*	*b*	*b*	*b*
g	*a*	*a*	*c*	*c*	*a*	*a*	*c*	*c*
h	*b*	*d*	*b*	*d*	*b*	*d*	*b*	*d*

The desirability of a consequence neither depends on the act nor on the state of nature leading to it.

In constructing a table like Table 1 some of the states of nature may be deleted if the decision-maker is certain that they cannot occur.

The next step in the process of selecting the best act is to construct 'conceivable' acts, which are not feasible. Thus the set of acts, *F*, in Savage's set-up consists of all functions from the set of states of nature, *S*, to the set *C* of consequences. In our example there are 4^8 acts. Of these, three acts *f, g* and *h* are actually feasible: The additional 65533 acts are only conceivable. The construction of the conceivable acts and the possibility of ranking all acts of *F* is a basic assumption of the present approach. For the sake of presentation we will in the next subsection assume the validity of the expected utility theory and then we will return to the rationale of our construction.

Suppose for the present that the decision-maker, in choosing between acts, indeed computes the expected utility of each act, and selects a feasible act with the highest expected utility. Thus we are assuming that he has assigned probability $P(s)$ to every state of nature s in S, and the utility $U(c)$ to every consequence c in C. So, given an act f in F, the expected utility $EU(f)$ of f equals $\sum_{s \in S} P(s)U[f(s)]$. More generally, if the set S is infinite, then P is a finitely additive probability measure defined on all events (i.e. subsets of S), and $EU(f)$ equals $\int U[f(s)]\, dP(s)$ (assuming the integral to exist; say U is bounded; see the Appendix, on Mathematical Expectation, section "Appendix: Mathematical Expectation"). So in fact in this case the decision-maker has a well-defined 'preference relation' (i.e. binary relation) $\geq$ on the set of acts *F*, with, for all *f, g* in *F*:

$$f \gtrsim g \text{ iff } \mathrm{EU}(f) \geq EU(g) \tag{1}$$

It is easily seen that the preference relation, defined in (1), is not affected when the utility function $U \rightarrow C\ R$ is replaced by any positive linear transformation of it (say $(\bar{U}: c \rightarrow \alpha U(c) + \beta$, for some real β and positive α).

If a preference relation, $\gtrsim$, over acts is derived from comparisons of expected utility as in (1), then it must satisfy several properties. We follow the terminology and order of Savage (1954). He listed seven postulates, five of which (P1 up to P4, and P7) are implied by (1). Postulate P1 says that the preference relation is complete ($f \gtrsim g$ or $g \gtrsim f$ for all acts *f, g*) and transitive. Postulate P2 is referred to as the sure-thing principle. It says that, when comparing two acts, only those states of nature matter, on which these acts differ. In other words, for the comparison between two acts, if they coincide on an event A, it really does not matter what actually the consequence is for each state in A. Thus P2 makes it possible to derive a preference relation over acts, conditioned on the event A^c; this for any event A.

Postulate P3 entails that the desirability of a consequence does not depend on the combination of state and act that lead to it; hence the possibility to express the desirability of consequences by a utility function on *C*.

P4 guarantees that the preference relation over acts induces a qualitative probability relation ('at least as probable as') over events, which is transitive and complete. P7 is a technical monotonicity condition.

P5 and P6 are Savage's only postulates which are not a necessary implication of (1). P5 simply serves to exclude the trivial case where the decision-maker is indifferent between any two acts. P6 implies some sort of continuity of the preference relation, and non-atomicity of the probability measure; the last term means that any non-impossible event can be partitioned into two non-impossible events. Hence there must be an infinite number of states.

Savage's great achievement was not to *assume* (1), but to show that his list of postulates P1–P7 *implies* that the preference relation over acts has

an expected utility representation as in (1). Savage argued compellingly for the appropriateness of his postulates. Furthermore, Savage showed that the probability measure in (1) is uniquely determined by the preference relation $\geq$, and that the utility function is unique up to a positive linear transformation.

The significance of Savage's achievement is that it gives the first, and until today most complete, conceptual foundation to expected utility. Savage's conclusion, to use expected utility for the selection of optimal acts, can be used even if we do not have the structure and the seven postulates of Savage. Indeed, the assumption needed on consequences, states, acts, and preferences, is that they can be extended so as to satisfy all requirements of Savage's model. Also other models, as mentioned in section "Other Approaches and Bibliographical Remarks", can be used to obtain expected utility representations.

Given a decision problem under uncertainty, if we assume that it can be embedded in Savage's framework, then it is not necessary to actually carry out this embedding. In other words, if the decision-maker is convinced that in principle it is possible to construct the conceivable acts as in subsection and the ranking of all acts in accordance with the postulates, then this construction does not have to be made. Instead one can directly try to assess probabilities and utilities, and apply the expected utility criterion. As an example, suppose a market-vendor has to decide whether to order 50 portions of ice-cream (f), or not (g). One portion costs \$1, and is sold for \$2. If the weather will be nice the next day, the school nearby will allow the children to go to the market, and all 50 portions, if ordered, will be sold, yielding a profit of \$50. If the weather is not nice, no portion will be sold. We assume that the ice-cream cannot be kept in stock and hence bad weather will yield a 'gain' of $\$ - 50$ if the portions have been ordered.

Instead of embedding the above example into Savage's framework, the market salesman may immediately assess P_1 (or $1 - P_1$), the probability for good (bad) weather; next assess the utilities of gaining \$50, \$0 and –\$50; finally order the 50 portions if $P_1 U(\$50) + (1 - P_1)\, U(-\$50) > U(\$0)$.

Theoretical conclusions can be derived from the mere assumption of expected utility maximization, without an actual assessment of the probabilities and utilities. Examples are the theories of attitudes towards risk, with applications to insurance, portfolio choice, etc. The validity of these applications depends on expected utility theory, which in turn depends on the plausibility of Savage's model (or other derivations of expected utility).

Another important theoretical application of Savage's model is to neo-Bayesian statistics. For applied statistics, in this vein, the availability of a 'prior distribution,' as proved by Savage's approach, is essential.

Expected Utility when Applied to Decisions Under Risks; The Von Neumann–Morgenstern Approach

Special and extreme cases of decisions under uncertainty are decisions in 'risky' situations. In decisions under uncertainty, as exposited in the previous section, the decision-maker who follows the dictum of expected utility has to assign utilities to the consequences and probabilities to the states. He can do it by mimicking the proof of Savage's theorem, or more directly by organizing his information, as the case may be.

Decision-making under risk considers the special case where the formulation of the problem for the decision-maker includes probabilities for the events, so that he only has to derive the utilities of consequences. As an example, consider a gambler in a casino who assumes that the roulette is really unbiased, so that each number has probability 1/37 (or 1/38). Another example is the St Petersburg paradox, described in subsection.

Within the framework of expected utility theory, for the evaluation of an act, only its probability distribution over the consequences has to be taken into account. Thus, for decision-making under risk, with probabilities known in advance, one may just as well describe acts as probability distributions over consequences instead of as functions from the states to the consequences.

Let us denote by L the set of probability distributions over C with finite support. We refer to them as lotteries. Von Neumann and Morgenstern (1947, Appendix) suggested conditions on a preference relation $\geq$ between lotteries, necessary and sufficient for the existence of a real-valued utility function U on C, such that for any two lotteries P and Q in L:

$$P \geq Q \text{ iff } \sum_{c \in C} P(c)U(c) \geq \sum_{c \in C} Q(c)U(c) \quad (2)$$

It is easy to see that the utility function, U, is unique up to positive linear transformations. Before we present a version of von Neumann-Morgenstern's theorem, recall that for any $0 \leq \alpha \leq 1$, and for any two lotteries P and W, $R := \alpha P + (1 - \alpha)Q$ is again a lottery, assigning probability $R(c) = \alpha P(c) + (1 - \alpha)Q(c)$ to any c in C. Also note that the assumption that all lotteries are given, is sometimes as heroic as Savage's assumption that all functions from S to C are conceivable acts.

The first axiom of von Neumann-Morgenstern, NM1, says that the preference relation over the lotteries is complete and transitive. NM2, the continuity axiom, says that, if $P \succsim Q \succ R$, then there are α, β in]0, 1[, such that $\alpha R + (1 - \alpha)P \succ Q \succ \beta P + (1 - \beta)R$. Here the strict preference relation $\succ$ is derived from $\succsim$ in the usual way: $P \succ Q$ if $P \succsim Q$ and not $Q \succsim P$.

The third axiom NM3 is the independence axiom. It says that for α in]0,1], P is preferred to Q iff $\alpha P + (1 - \alpha)R$ is preferred to $\alpha Q + (1 - \alpha)R$. This condition is the antecedent of Savage's sure-thing principle, and is the most important innovation of the above axioms.

Von Neumann and Morgenstern originally stated their theorem for more general sets than L. They did it for so-called mixture spaces, i.e. spaces endowed with some sort of convex combination operation. This has been done more precisely by Herstein and Milnor (1953).

Von Neumann and Morgenstern introduced their theory of decision-making under risk as a normative tool for playing zero-sum games in strategic form. There the 'player' (i.e. decision-maker) can actually construct any lottery he wishes over his pure-strategies (but not over his consequences).

The theorem of von Neumann and Morgenstern, stated above, is a major step in the proof of Savage's theorem.

Recently there has been much research on decision making under risk for its own end. Some of this research is experimental, subjects are asked to express their preferences between lotteries. These experiments, or polls, reveal violations of most of the axioms. They lead to representations different from expected utility.

Other Approaches and Bibliographical Remarks

The first suggestion for expected utility theory in decision-making under uncertainty in the vein of Savage was Ramsey's (1931). His model was not completely formalized. The work of Savage was influenced by de Finetti's approach to probabilities, as in de Finetti (1931, 1937). The decision theoretic framework to which Savage's expected utility model owes much is that of Wald (1951), who regards a statistician as a decision-maker.

A model which can be considered intermediate between those of Savage and von Neumann and Morgenstern is that considered by Anscombe and Aumann (1963). Formally it is a special case of a mixture set, but like Savage it introduces states of nature, and gives a simultaneous derivation of probabilities for the states, and of utilities for the consequences. A consequence in this model consists of a lottery over deterministic outcomes; this involves probabilities known in advance, as in the approach of von Neumann and Morgenstern. The Anscombe and Aumann theory, as well as most of the technical results up to 1970, are presented in detail in Fishburn (1970).

In the expected utility theory, described above, the desirability (utility) of consequences does not depend on acts or states of nature. This is a restriction in many applications. For example the desirability of family income may depend on whether the state of nature is 'head of family alive' or 'head of family deceased'. Karni (1985) summarized and developed the expected utility theory

without the restrictive assumption of state-independent preferences over consequences.

Ellsberg (1961) argued against the expected utility approach of Savage by proposing an example, inconsistent with it. A way of resolving the inconsistency is to relax the additivity property of the involved probability measures. Schmeidler (1984) formulated expected utility theory with non-additive probabilities for the framework of Anscombe and Aumann (1963). Gilboa (1985) did the same for the original framework of Savage. Wakker (1986) obtained expected utility representation, including the non-additive case, for a finite number of states of nature and non-linear utility.

Appendix: Mathematical Expectation

Expectation with respect to finitely additive probability. A non-empty collection $\sum$ of subsets (called events) of a non-empty set S is said to be an algebra if it contains the complement of each set belonging to it, and it contains the union of any two sets belonging to it. A (finitely additive) probability P on $\sum$ assigns to every event in $\sum$ a number between 0 and 1 such that $P(S) = 1$ and for any two *disjoint events* A and B, $P(A \cup B) = P(A) + P(B)$.

A random variable X is a real-valued function on S such that, for any open or closed (bounded or unbounded) interval I, $\{s \in S \mid X(s) \in I\}$ (or $[X \in I]$ for short) is an event i.e., in Σ. Given such a random variable X, its (mathematical) expectation is:

$$E(X) = \int_0^\infty P[X \geq \alpha]\,\mathrm{d}\alpha - \int_{-\infty}^0 (1 - P[X \geq \alpha])\,\mathrm{d}\alpha \qquad (3)$$

where the integration above is Riemann-integration and it is assumed that the integral exist. The integrands in (3) are monotonic, so $E(X)$ exist *if* X is bounded. If the random variable X has finitely many values, say $x_1, \ldots, x_n$ then (3) reduces to

$$E(X) = \sum_{i=1}^{n} P(X = x_i)x_i, \qquad (4)$$

However, an equation like that above may not hold if the random variable obtains countably many different values. An example will be provided in subsection.

σ-additive probability. Kolmogorov (1933) imposed an additional continuity assumption on probability P on $\sum$: To simplify presentation he first assumed that Σ is a σ-algebra, i.e., an algebra such that for every sequence of events $(A_i)_{i=1}^{\infty}$ it contains its union $\mathrm{U}_{i=1}^{\infty} A_i$. He then required that $P\left(\mathrm{U}_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ if the A_i's are pairwise disjoint.

This last property is referred to as σ -additivity of the probability P. In this way Kolmogorov transformed large parts of probability theory into (a special case of) measure theory. Thus an expectation of a random variable X is

$$E(X) = \int_s X(s)\mathrm{d}P(s) \qquad (5)$$

where the right side is a Lebesgue integral (if it exists. . .), defined as a limit of integrals of random variables with countably many values. Let Y be such a random variable with values $(y_i)_{i=1}^{\infty}$, then

$$E(Y) = \sum_{i=1}^{\infty} P(Y = y_i)y_i \qquad (6)$$

if the right side is absolutely convergent.

An example will now be introduced of a finitely additive probability, i.e. a probability for which (4) holds but (6) does not hold. Let S be the set of rational numbers in the interval [0, 1] and let Σ be the algebra of all subsets of S. (It is in fact a σ-algebra.) For $0 \leq \alpha \leq \beta \leq 1$ define $P(S \cap [\alpha,\beta]) = \beta - \alpha$ and extend P to all subsets of Σ. For each s in S, $P(s) = 0$. Since S is countable we can write $S = \{s_1, s_2, \ldots\}$ and $1 = P(S) > \Sigma_{i=1}^{\infty} P(si) = 0$. Defining $Y(s_i) = 1/i$ for all i, we get a contradiction to (6). The finitely additive probability P has also the property implied by Savage's P6 (see 2.4): If $P(A) > 0$ then there is an event $B \subset A$ such that $0 < P(B) < \mathrm{P(A)}$.

Distributions. A non-decreasing right continuous function on the extended real line is called a distribution function if $F(-\infty) = 0$ and $F(\infty) = 1$. Given a random variable X, its distribution function F_x is defined by $F_x(\alpha) = P(X \leqq \alpha)$ for all real α. Then

$$E(X) = \int_0^\infty [1 - F_X(\alpha)]\mathrm{d}\alpha - \int_{-\infty}^0 F(\alpha)\mathrm{d}\alpha \quad (7)$$

which is the dual of formula (3). If the distribution F_x is smooth we say that the random variable X has a density f_x: $R \to R$, which is the derivative of F_x. In this case

$$E(X) = \int_{-\infty}^\infty = \alpha f(\alpha)\mathrm{d}\alpha \quad (8)$$

Non-additive probability. A function P: $\Sigma \to [0, 1]$ is said to be *non-additive* probability (or capacity) if $P(S) = 1$, $P(\phi) = 0$ *and for* $A \subset B$, $P(A) \leq P(B)$. Choquet (1954) suggested to integrate a random variable with respect to non-additive probability by formula (3).

See Also

- Allais paradox
- Mean Value
- Risk
- Subjective probability
- Uncertainty
- Utility Theory and Decision Theory

Bibliography

Anscombe, F.J., and R.J. Aumann. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34: 199–205.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5, 175–192. Translated into English by L. Sommer (1954) as: Exposition of a new theory on the measurement of risk, *Econometrica* 12, 23–36; or in *Utility theory: A book of readings*, ed. A.N. Page. New York: Wiley, 1986.

Choquet, G. 1953–54. Theory of capacities. *Annales de l'Institut Fourier* (Grenoble), 131–295.

de Finetti, B. 1931. Sul significato soggettivo della probabilita. *Fundamenta Mathematicae* 17: 298–329.

de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* 7, 1–68. Translated into English in *Studies in Subjective Probability*, ed. H.E. Kyburg and H.E. Smokler, 1964. New York: Wiley.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Fishburn, P.C. 1970. *Utility theory for decision making*. New York: Wiley.

Gilboa, I. 1986. Non-additive probability measures and their applications in expected utility theory. PhD thesis submitted to Tel Aviv University.

Herstein, I.N., and J. Milnor. 1953. An axiomatic approach to measurable utility. *Econometrica* 21: 291–297.

Karni, E. 1985. *Decision-making under uncertainty: The case of state-dependent preferences*. Cambridge, Mass.: Harvard University Press.

Kolmogorov, A.N. 1933. *Grundbegriffe der Wahrscheinlichkeitrechnung*. Berlin. Translated into English by Nathan Morrison (1950, 2nd edn, 1956). New York: Chelsea Publishing Company.

Loeve, M. 1963. *Probability theory*, 3rd ed. Princeton: Van Nostrand.

von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press.

Ramsey, F.P. 1931. Truth and probability. In *The foundations of mathematics and other logical essays*, ed. R.B. Braithwaite. New York: Harcourt, Brace.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley, 2nd edn, 1972.

Schmeidler, D. 1984. Subjective probability and expected utility without additivity. CARESS, University of Pennsylvania and IMA University of Minnesota, mimeo.

Wald, A. 1951. *Statistical decision functions*. New York: Wiley.

Wakker, P.P. 1986. Representations of choice situations. PhD thesis, University of Tilburg, Department of Economics.

Expected Utility Hypothesis

Mark J. Machina

Abstract

The expected utility hypothesis – that is, the hypothesis that individuals evaluate uncertain prospects according to their expected level of 'satisfaction' or 'utility' – is the predominant

descriptive and normative model of choice under uncertainty in economics. It provides the analytical underpinnings for the economic theory of risk-bearing, including its applications to insurance and financial decisions, and has been formally axiomatized under conditions of both objective (probabilistic) and subjective (event-based) uncertainty. In spite of evidence that individuals may systematically depart from its predictions, and the development of alternative models, expected utility remains the leading model of economic choice under uncertainty.

Keywords

Arrow–Pratt index of absolute risk aversion; Bernoulli, D.; Bernoulli, N.; Cobb–Douglas functions; Comparative likelihood; Consumer theory; Cramer, G.; Environmental economics; Expected utility hypothesis; First-order stochastic dominance preference; Increasing risk; Independence axiom; Inequality (measurement); International trade; Lotteries; Malinvaud, E.; Marschak, J.; Menger, K.; Objective vs. subjective uncertainty; Ordinal revolution; Preference functions; Preference orderings; Probability; Risk; Risk aversion; St Petersburg paradox; Stochastic dominance; Subjective probability; Sure-thing principle; Transitivity; Uncertainty; von Neumann–Morgenstern utility function

JEL Classifications

D8

The expected utility hypothesis is the predominant descriptive and prescriptive theory of individual choice under conditions of risk or uncertainty.

The expected utility hypothesis of behaviour towards risk is the hypothesis that the individual possesses (or acts as if possessing) a 'von Neumann–Morgenstern utility function' $U(\cdot)$ or 'von Neumann–Morgenstern utility index' $\{U_i\}$ defined over some set $\mathscr{X}$ of alternative possible outcomes, and when faced with alternative risky prospects or 'lotteries' over these outcomes, will choose the prospect that maximizes the expected value of $U(\cdot)$ or $\{U_i\}$. Since the outcomes could be alternative wealth levels, multidimensional commodity bundles, time streams of consumption, or even non-numerical consequences (such as a trip to Paris), this approach can be applied to a tremendous variety of situations, and most theoretical research in the economics of uncertainty, as well as virtually all applied work in the field (for example, insurance or investment decisions) is undertaken in the expected utility framework.

As a branch of modern consumer theory (for example, Debreu 1959, ch. 4), the expected utility model proceeds by specifying a set of objects of choice and assuming that the individual possesses a preference ordering over these objects which may be represented by a real-valued maximand or 'preference function' $V(\cdot)$, in the sense that one object is preferred to another if and only if it is assigned a higher value by this preference function. However, the expected utility model differs from the theory of choice over non-stochastic commodity bundles in two important respects. The first is that, since it is a theory of choice under uncertainty, the objects of choice are not deterministic outcomes but rather uncertain prospects. The second difference is that, unlike in the non-stochastic case, the expected utility model imposes a very specific restriction on the functional form of the preference function $V(\cdot)$.

The formal representation of the objects of choice, and hence of the expected utility preference function, depends upon the set of possible outcomes. When the outcome set $\mathscr{X} = \{x_1, \ldots, x_n\}$ is finite, we can represent any probability distribution over this set by its vector of probabilities $\boldsymbol{P} = (p_1, \ldots, p_n)$ (where $p_i = \text{prob}(x_i)$) and the expected utility preference function takes the form

$$V(\boldsymbol{P}) = V(p_1, \ldots, p_n) \equiv \sum U_i p_i.$$

When the outcome set consists of the real line or some interval subset of it, probability distributions can be represented by their cumulative distribution functions $F(\cdot)$ (where $F(x) = \text{prob}(\tilde{x} \leq x)$), and the expected utility preference function takes the form $V(F) \equiv \int U(x)\mathrm{d}F(x)$ (or $\int U(x)f(x)\mathrm{d}x$

when $F(\cdot)$ possesses a density function $f(\cdot)$). When the outcomes are commodity bundles of the form $(z_1, \ldots, z_m)$, cumulative distribution functions are multivariate, and the preference function takes the form $\int \ldots \int U(z_1, \ldots, z_m)\, dF(z_1, \ldots, z_m)$. The expected utility model derives its name from the fact that in each case the preference function consists of the mathematical expectation of the von Neumann–Morgenstern utility function $U(\cdot)$, $U(\cdot, \ldots, \cdot)$ or utility index $\{U_i\}$ with respect to the probability distribution $F(\cdot)$, $F(\cdot, \ldots, \cdot)$ or $\boldsymbol{P}$.

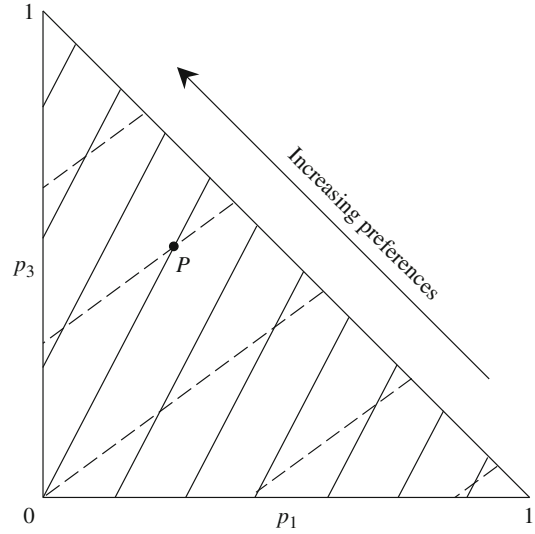
Mathematically, the hypothesis that the preference function $V(\cdot)$ takes the form of a statistical expectation is equivalent to the condition that it be 'linear in the probabilities', that is, either a weighted sum of the components of $\boldsymbol{P}$ (i.e. $\sum U_i p_i$) or else a weighted integral of the functions $F(\cdot)$ or $f(\cdot)$ ($\int U(x)\mathrm{d}F(x)$ or $\int U(x)f(x)\mathrm{d}x$). Although this still allows for a wide variety of attitudes towards risk depending upon the shape of the utility function $U(\cdot)$ or utility index $\{U_i\}$, the restriction that $V(\cdot)$ be linear in the probabilities is the primary empirical feature of the expected utility model, and provides the basis for many of its observable implications and predictions.

It is important to distinguish between the preference function $V(\cdot)$ and the von Neumann–Morgenstern utility function $U(\cdot)$ (or index $\{U_i\}$) of an expected utility maximizer, in particular with regard to the prevalent though mistaken belief that expected utility preferences are somehow 'cardinal' in a sense not exhibited by preferences over non-stochastic commodity bundles. As with any real-valued representation of a preference ordering, an expected utility *preference function* $V(\cdot)$ is 'ordinal' in that it may be subject to any increasing transformation without affecting the validity of the representation – thus, the preference functions $\int U(x)\mathrm{d}F(x)$ and $[\int U(x)dF(x)]^3$ represent identical risk preferences. On the other hand, the von Neumann–Morgenstern utility function $U(\cdot)$ is 'cardinal' in the sense that a different utility function $U^*(\cdot)$ will generate an ordinally equivalent preference function $V^*(F) \equiv \int U^*(x)\mathrm{d}F(x)$ if and only if it satisfies the cardinal relationship $U^*(x) \equiv a \cdot U(x) + b$ for some $a > 0$ (in which case $V^*(F) \equiv a \cdot V(F) + b$. However, the same distinction holds in the theory of preferences over non-stochastic commodity bundles: the Cobb–Douglas preference function $\alpha \cdot \ln(z_1) + \beta \cdot \ln(z_2) + \gamma \cdot \ln(z_3)$ (written here in its additive form) can be subject to any increasing transformation and is clearly ordinal, even though a vector of parameters $(\alpha^*, \beta^*, \gamma^*)$ will generate an ordinally equivalent additive form $\alpha^* \cdot \ln(z_1) + \beta^* \cdot \ln(z_2) + \gamma^* \cdot \ln(z_3)$ if and only if it satisfies the cardinal relationship $(\alpha^*, \beta^*, \gamma^*) = \lambda \cdot (\alpha, \beta, \gamma)$ for some $\lambda > 0$.

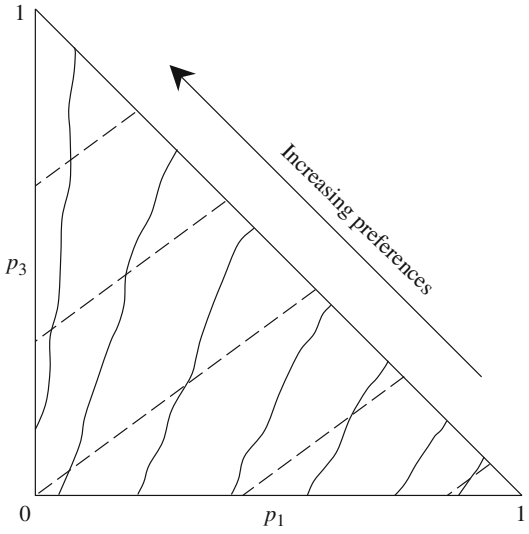
In the case of a simple outcome set of the form $\{x_1, x_2, x_3\}$, it is possible to graphically illustrate the 'linearity in the probabilities' property of expected utility preferences. Since every probability distribution (p_1, p_2, p_3) over these outcomes must satisfy $p_1 + p_2 + p_3 = 1$, we may represent such distributions by points in the unit triangle in the (p_1, p_3) plane, with p_2 given by $p_2 = 1 - p_1 - p_3$ (Figs. 1 and 2). Since they represent the loci of solutions to the equations

$$\begin{aligned} U_1p_1 + U_2p_2 + U_3p_3 &= U_2 - [U_2 - U_1] \cdot p_1 \\ &\quad + [U_3 - U_2] \cdot p_3 \\ &= \text{constant} \end{aligned}$$

for the fixed utility indices $\{U_1, U_2, U_3\}$, the indifference curves of an expected utility maximizer consist of parallel straight lines in the triangle,



Expected Utility Hypothesis, Fig. 1 Expected utility indifference curves



Expected Utility Hypothesis, Fig. 2 Non-expected utility indifference curves

with slope $[U_2 - U_1]/[U_3 - U_2]$, as illustrated by the solid lines in Fig. 1. Indifference curves which do *not* satisfy the expected utility hypothesis (that is, are not linear in the probabilities) are illustrated by the solid curves in Fig. 2.

When the outcomes consist of different wealth levels $x_1 < x_2 < x_3$, this diagram can be used to illustrate other possible features of an expected utility maximizer's attitudes towards risk. On the principle that more wealth is better, it is typically postulated that any change in a distribution (p_1, p_2, p_3) which increases p_3 at the expense of p_2, increases p_2 at the expense of p_1, or both, will be preferred: this property is known as 'first-order stochastic dominance preference'. Since such shifts of probability mass are represented by north, west, or north-west movements in the diagram, first-order stochastic dominance preference is equivalent to the condition that indifference curves are upward sloping, with more preferred indifference curves lying to the north-west. Algebraically, this is equivalent to the condition $U_1 < U_2 < U_3$.

Another widely (though not universally) hypothesized aspect of attitudes towards risk is that of 'risk aversion' (for example, Arrow 1974, ch. 3; Pratt 1964). To illustrate this property, consider the dashed lines in Fig. 1, which represent loci of solutions to the equations

$$\begin{aligned} x_1p_1 + x_2p_2 + x_3p_3 &= x_2 - [x_2 - x_1] \cdot p_1 \\ &\quad + [x_3 - x_2] \cdot p_3 \\ &= \text{constant} \end{aligned}$$

and hence may be termed 'iso-expected *value* loci'. Since north-east movements along any of these loci consist of increasing the tail probabilities p_1 and p_3 at the expense of the middle probability p_2 in a manner which preserves the mean of the distribution, they correspond to what are termed 'mean-preserving increases in risk' (Rothschild and Stiglitz 1970, 1971). An individual is said to be 'risk averse' if such increases in risk always lead to less preferred indifference curves, which is equivalent to the graphical condition that the indifference curves be steeper than the iso-expected value loci. Since the slope of the latter is given by $[x_2 - x_1]/[x_3 - x_2]$, this is equivalent to the algebraic condition that $[U_2 - U_1]/[x_2 - x_1] > [U_3 - U_2]/[x_3 - x_2]$. Conversely, individuals who *prefer* mean-preserving increases in risk are termed 'risk loving': such individuals' indifference curves will be flatter than the iso-expected value loci, and their utility indices will satisfy $[U_2 - U_1]/[x_2 - x_1] < [U_3 - U_2]/[x_3 - x_2]$.

Note finally that the indifference map in Fig. 1 indicates that the lottery $\boldsymbol{P}$ is indifferent to the origin, which represents the degenerate lottery yielding x_2 with certainty. In such a case the amount x_2 is said to be the 'certainty equivalent' of the lottery $\boldsymbol{P}$. The fact that the origin lies on a lower iso-expected value locus than $\boldsymbol{P}$ reflects a general property of risk-averse preferences, namely, that the certainty equivalent of any lottery will always be less than its mean. (For risk lovers, the opposite is the case.)

When the outcomes are elements of the real line, it is possible to represent the above (as well as other) aspects of preferences in terms of the shape of the von Neumann–Morgenstern utility function $U(\cdot)$, as seen in Figs. 3 and 4. In each figure, consider the lottery which assigns the probabilities 2/3:1/3 to the outcome levels x': x''. The expected value of this lottery, $\overline{x} = 2/3 \cdot x' + 1/3$

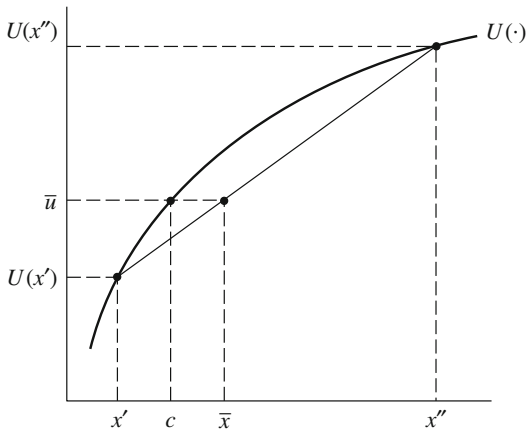
$\cdot x''$, lies between these two values, two-thirds of the way towards x'. The expected *utility* of this lottery, $\overline{u} = 2/3 \cdot U(x') + 1/3 \cdot U(x'')$ lies between $U(x')$ and $U(x'')$ on the vertical axis, two-thirds of the way towards $U(x')$. The point $(\overline{x},\overline{u})$ thus lies on the line segment connecting the points $(x', U(x'))$ and $(x'', U(x''))$, two-thirds of the way towards the former. In each figure, the certainty equivalent of this lottery is given by the sure outcome c that also yields a utility level of $\overline{u}$.

The property of first-order stochastic dominance preference can be extended to the case of distributions over the real line (Quirk and Saposnick 1962), and it is equivalent to the condition that $U(x)$ be an increasing function of x, as in Figs. 3 and 4. It is also possible to generalize the notion of a mean-preserving increase in risk to density functions or cumulative distribution functions (Rothschild and Stiglitz 1970, 1971), and the earlier algebraic condition for risk aversion generalizes to the condition that $U''(x) < 0$ for all x, that is, that the von Neumann–Morgenstern utility function $U(\cdot)$ be concave, as in Fig. 3. As before, risk aversion implies that the certainty equivalent of any lottery will lie below its mean, as seen in Fig. 3; the opposite is true for the convex utility function of a risk lover, as in Fig. 4. Two of the earliest and most important analyses of risk attitudes in terms of the shape of the von Neumann–Morgenstern utility function are those of Friedman and Savage (1948) and Markowitz (1952).

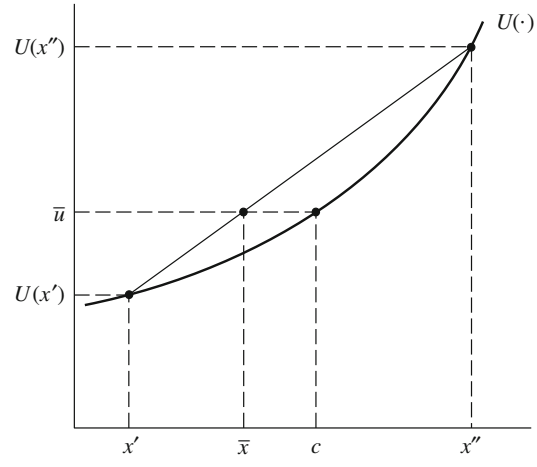
Analytics

The tremendous analytic capabilities of the expected utility model derive largely from the work of Arrow (1974) and Pratt (1964), who showed that the 'degree' of concavity of the utility function provides a measure of an expected utility maximizer's 'degree' of risk aversion. Formally, the Arrow–Pratt characterization of comparative risk aversion is the result that the following conditions on a pair of (increasing, twice differentiable) von Neumann–Morgenstern utility functions $U_a(\cdot)$ and $U_b(\cdot)$ are equivalent:

- $U_a(\cdot)$ is a concave transformation of $U_b(\cdot)$ (that is, $U_a(x) \equiv \rho(U_b(x))$ for some increasing concave function $\rho(\cdot)$),
- $-U_a''(x)/U_a'(x) \geq -U_b''(x)/U_b'(x)$ for each x,
- if c_a and c_b solve $U_a(c_a) = \int U_a(x)\mathrm{d}F(x)$ and $U_b(c_b) = \int U_b(x)\mathrm{d}F(x)$ for some distribution $F(\cdot)$, then $c_a \leq c_b$,



Expected Utility Hypothesis, Fig. 3 Von Neumann–Morgenstern utility function of a risk averse individual



Expected Utility Hypothesis, Fig. 4 Von Neumann–Morgenstern utility function of a risk loving individual

and if $U_a(\cdot)$ and $U_b(\cdot)$ are both concave, these conditions are in turn equivalent to:

- if $r > 0$, $\mathrm{E}[\tilde{z}] > r$, $\mathrm{prob}(\tilde{z} < r) > 0$, and α_a and α_b maximize $\int U_a((I - \alpha) \cdot r + \alpha \cdot z)\mathrm{d}F(z)$ and $\int U_b((I - \alpha) \cdot r + \alpha \cdot z)\mathrm{d}F(z)$ respectively, then $\alpha_a \leq \alpha_b$.

The first two of these conditions provide equivalent formulations of the notion that $U_a(\cdot)$ is a more concave function than $U_b(\cdot)$. The curvature measure $R(x) \equiv -U''(x)/U'(x)$ is known as the 'Arrow–Pratt index of (absolute) risk aversion', and plays a key role in the analytics of the expected utility model. The third condition states that the more risk averse utility function $U_a(\cdot)$ will never assign a higher certainty equivalent to any lottery $F(\cdot)$ than will $U_b(\cdot)$. The final condition pertains to the individuals' respective demands for risky assets. Specifically, assume that each must allocate $\$I$ between two assets, one yielding a riskless (gross) return of r per dollar, and the other yielding a risky return $\tilde{z}$ with a higher expected value but with some chance of doing worse than r. This condition says that the less risk-averse utility function $U_b(\cdot)$ will generate at least as great a demand for the risky asset as the more risk-averse utility function $U_a(\cdot)$. It is important to note that it is the *equivalence* of the above concavity, certainty equivalent and asset demand conditions which makes the Arrow-Pratt characterization such an important result in expected utility theory. (Ross 1981, provides an alternative, stronger, characterization of comparative risk aversion.)

Although the applications of the expected utility model extend to virtually all branches of economic theory (for example, Hey 1979), much of the flavour of these analyses can be gleaned from Arrow's (1974, ch. 3) analysis of the portfolio problem of the previous paragraph. If we rewrite $(I - \alpha) \cdot r + \alpha \cdot z$ as $I \cdot r + \alpha \cdot (z - r)$, the first-order condition for this problem can be expressed as:

$$\int z \cdot U'(I \cdot r + \alpha \cdot (z - r))\mathrm{d}F(z) - \int r \cdot U'(I \cdot r + \alpha \cdot (z - r))\mathrm{d}F(z) = 0,$$

that is, the marginal *expected* utility of the last dollar allocated to each asset is the same. The second-order condition can be written as:

$$\int (z - r)^2 \cdot U'''(I \cdot r + \alpha \cdot (z - r))\ \mathrm{d}F(z) < 0$$

and is ensured by the property of risk aversion (i.e. $U''(\cdot) < 0$).

As usual, we may differentiate the first-order condition to obtain the effect of a change in some parameter, say initial wealth I, on the optimal level of investment in the risky asset (the optimal value of α). Differentiating the first-order condition (including α) with respect to I, solving for $d\alpha/dI$, and invoking the second-order condition and the positivity of r yields that this derivative possesses the same sign as:

$$\int (z - r) \cdot U''(I \cdot r + \alpha \cdot (z - r))\ \mathrm{d}F(z).$$

Substituting $U''(x) \equiv -R(x) \cdot U'(x)$ and subtracting $R(I \cdot r)$ times the first-order condition yields that this expression is equal to:

$$-\int (z - r) \cdot [R(I \cdot r + \alpha \cdot (z - r)) - R(I \cdot r)] \cdot U'(I \cdot r + \alpha \cdot (z - r))\mathrm{d}F(z).$$

On the assumption that α is positive and $R(\cdot)$ is monotonic, the expression $(z - r) \cdot [R(I \cdot r + \alpha \cdot (z - r)) - R(I \cdot r)]$ will possess the same sign as $R'(\cdot)$. This implies that the derivative $\mathrm{d}\alpha/\mathrm{d}I$ will be positive (negative) whenever the Arrow–Pratt index $R(x)$ is a decreasing (increasing) function of the individual's wealth level x. In other words, an increase in initial wealth will always increase (decrease) demand for the risky asset if and only if $U(\cdot)$ exhibits decreasing (increasing) absolute risk aversion in wealth. Further examples of the analytics of the expected utility model may be found in the above references, as well as the surveys of Hirshleifer and Riley (1979), Lippman and McCall (1981), Machina (1983) and Karni and Schmeidler (1991).

Axiomatic Development

Although there exist dozens of formal axiomatizations of the expected utility model, most proceed by specifying an outcome space and postulating that the individual's preferences over probability distributions on this outcome space satisfy the following four axioms: completeness, transitivity, continuity and the Independence Axiom. Although it is beyond the scope of this entry to provide a rigorous derivation of the expected utility model in its most general setting, it is possible to illustrate the meaning of the axioms and sketch a proof of the expected utility representation theorem in the simple case of a finite outcome set $\{x_1, \ldots, x_n\}$.

Recall that in such a case the objects of choice consist of probability distributions $\boldsymbol{P} = (p_1, \ldots, p_n)$ over $\{x_1, \ldots, x_n\}$, so that the following axioms refer to the individuals' weak preference relation $\succcurlyeq$ over these prospects, where $\boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$ is read '$\boldsymbol{P}^*$ is weakly preferred (that is, preferred or indifferent) to $\boldsymbol{P}$' (the associated strict preference relation $\succ$ and indifference relation $\sim$ are defined in the usual manner):

- *Completeness*: For any two distributions $\boldsymbol{P}$ and $\boldsymbol{P}^*$, either $\boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$, $\boldsymbol{P} \succcurlyeq \boldsymbol{P}^*$, or both.
- *Transitivity*: If $\boldsymbol{P}^{**} \succcurlyeq \boldsymbol{P}^*$ and $\boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$, then $\boldsymbol{P}^{**} \succcurlyeq \boldsymbol{P}$.
- *Mixture continuity*: If $\boldsymbol{P}^{**} \succcurlyeq \boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$, then there exists some $\lambda \in [0, 1]$ such that $\boldsymbol{P}^* \sim \lambda \cdot \boldsymbol{P}^{**} + (1 - \lambda) \cdot \boldsymbol{P}$.
- *Independence*: For any two distributions $\boldsymbol{P}$ and $\boldsymbol{P}^*$, $\boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$ if and only if $\lambda \cdot \boldsymbol{P}^* + (1 - \lambda) \cdot \boldsymbol{P}^{**} \succcurlyeq \lambda \cdot \boldsymbol{P} + (1 - \lambda) \cdot \boldsymbol{P}^{**}$ for all $\lambda \in [0, 1]$ and all $\boldsymbol{P}^{**}$

where $\lambda \cdot \boldsymbol{P} + (1 - \lambda) \cdot \boldsymbol{P}^{**}$ denotes the $\lambda : (1 - \lambda)$ 'probability mixture' of $\boldsymbol{P}$ and $\boldsymbol{P}^{**}$, that is, the lottery with probabilities $(\lambda \cdot p_1 + (1 - \lambda) \cdot p_1^{**}, \ldots, \lambda \cdot p_n + (1 - \lambda) \cdot p_n^{**})$.

The completeness and transitivity axioms are analogous to their counterparts in standard consumer theory. Mixture continuity states that if the lottery P^{**} is weakly preferred to P^* and P^* is weakly preferred to P, then exists some probability mixture of the most and least preferred lotteries which is indifferent to the intermediate one.

As in standard consumer theory, completeness, transitivity and continuity serve to establish the existence of a real-valued preference function $V(p_1, \ldots, p_n)$ which represents the relation $\succcurlyeq$, in the sense that $\boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$ if and only if $V(p_1^*, \ldots, p_n^*) \geq V(p_1, \ldots, p_n)$. It is the Independence Axiom which gives the theory its primary empirical content by implying that $\succcurlyeq$ can be represented by a linear preference function of the form $V(p_1, \ldots, p_n) \equiv \sum U_i p_i$. To see the meaning of this axiom, assume that individuals are always indifferent between a two-stage compound lottery and its probabilistically equivalent single-stage lottery, and that $\boldsymbol{P}^*$ happens to be weakly preferred to $\boldsymbol{P}$. In that case, the choice between the mixtures $\lambda \cdot \boldsymbol{P}^* + (1 - \lambda) \cdot \boldsymbol{P}^{**}$ and $\lambda \cdot \boldsymbol{P} + (1 - \lambda) \cdot \boldsymbol{P}^{**}$ is equivalent to being presented with a coin that has a $(1 - \lambda)$ chance of landing tails (in which case the prize will be $\boldsymbol{P}^{**}$) and being asked *before the flip* whether one would rather win $\boldsymbol{P}^*$ or $\boldsymbol{P}$ in the event of a head. The normative argument for the Independence Axiom is that either the coin will land tails, in which case the choice won't have mattered, or it will land heads, in which case one is 'in effect' facing a choice between $\boldsymbol{P}^*$ and $\boldsymbol{P}$ and one 'ought' to have the same preferences as before. Note finally that the above statement of the axiom in terms of the weak preference relation $\succcurlyeq$ also implies its counterparts in terms of strict preference and indifference.

In the following sketch of the expected utility representation theorem, expressions such as '$x_i \succcurlyeq x_j$' should be read as saying that the individual weakly prefers the degenerate lottery yielding x_i with certainty to that yielding x_j with certainty, and '$\lambda \cdot x_i + (1 - \lambda) \cdot x_j$' will be used to denote the $\lambda : (1 - \lambda)$ probability mixture of these two degenerate lotteries.

The first step in the proof is to define the von Neumann–Morgenstern utility index $\{U_i\}$ and the expected utility preference function $V(\cdot)$. Without loss of generality, we may order the outcomes so that $x_n \succcurlyeq x_{n-1} \succcurlyeq \ldots \succcurlyeq x_2 \succcurlyeq x_1$. Since $x_n \succcurlyeq x_i \succcurlyeq x_1$ for each outcome x_i, mixture

continuity implies that there exist scalars $\{Ui\} \subset [0, 1]$ such that $x_i \sim U_i \cdot x_n + (1 - U_i) \cdot x_1$ for each i (which implies $U_1 = 0$ and $U_n = 1$). Given this, define $V(\boldsymbol{P}) = \sum U_i p_i$ for each $\boldsymbol{P}$.

The second step is to show that each lottery $\boldsymbol{P} = (\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n)$ is indifferent to the mixture $\lambda \cdot x_n + (1 - \lambda) \cdot x_1$ where $\lambda = \sum U_i p_i$. Since $(p_1,\ldots,p_n)$ can be written as the n-component probability mixture $p_1 \cdot x_1 + p_2 \cdot x_2 + \ldots + p_n \cdot x_n$, and each outcome x_i is indifferent to the mixture $U_i \cdot x_n + (1 - U_i) \cdot x_1$, an n-fold application of the Independence Axiom yields that $\boldsymbol{P} = (p_1,\ldots,p_n)$ is indifferent to the mixture

$$p_1 \cdot [U_1 \cdot x_n + (1 - U_1) \cdot x_1] + p_2 \cdot [U_2 \cdot x_n + (1 - U_2) \cdot x_1] + \ldots + p_n \cdot [U_n \cdot x_n + (1 - U_n) \cdot x_1],$$

which is equivalent to $\left(\sum_{i=1}^{n} U_i p_i\right) \cdot x_n + \left(1 - \sum_{i=1}^{n} U_i p_i\right) \cdot x_1$.

The third step is to demonstrate that a mixture $\lambda^* \cdot x_n + (1 - \lambda^*) \cdot x_1$ is weakly preferred to a mixture $\lambda \cdot x_n + (1 - \lambda) \cdot x_1$ if and only if $\lambda^* \geq \lambda$. This follows immediately from the Independence Axiom and the fact that $\lambda^* \geq \lambda$ implies that these two lotteries may be expressed as the respective mixtures $(\lambda^* - \lambda) \cdot x_n + (1 - \lambda^* + \lambda) \cdot \boldsymbol{Q}$ and $(\lambda^* - \lambda) \cdot x_1 + (1 - \lambda^* + \lambda) \cdot \boldsymbol{Q}$, where $\boldsymbol{Q}$ is defined as the lottery $(\lambda/(1 - \lambda^* + \lambda)) \cdot x_n + ((1 - \lambda^*)/(1 - \lambda^* + \lambda)) \cdot x_1$.

The completion of the proof is now simple. For any two distributions $\boldsymbol{P}^* = \left(p_1^*,\ldots,p_n^*\right)$ and $\boldsymbol{P} = (p_1, \ldots, p_n)$, transitivity and the second step imply that $\boldsymbol{P}^* \succcurlyeq \boldsymbol{P}$ if and only if

$$\left(\sum_{i=1}^{n} U_i p_i^*\right) \cdot x_n + \left(1 - \sum_{i=1}^{n} U_i p_i^*\right) \cdot x_1 \succcurlyeq \left(\sum_{i=1}^{n} U_i p_i\right) \cdot x_n + \left(1 - \sum_{i=1}^{n} U_i p_i\right) \cdot x_1,$$

which by the third step is equivalent to the condition $\sum U_i p_i^* \geq \sum U_i p_i$, or in other words, that $V(\boldsymbol{P}^*) \geq V(\boldsymbol{P})$.

As mentioned, the expected utility model has been axiomatized many times and in many contexts. The most comprehensive accounts of the axiomatics of the model are undoubtedly Fishburn (1982) and Kreps (1988).

Subjective Expected Utility

In addition to the above setting of 'objective' (that is, probabilistic) uncertainty, it is possible to define expected utility preferences under conditions of 'subjective' uncertainty. In this case, uncertainty is represented by a set $\mathscr{S}$ of mutually exclusive and exhaustive 'states of nature,' which can be a finite set $\{s_1,\ldots, s_n\}$ (as with a horse race), a real interval $[\underline{s}, \overline{s}] \subseteq R^1$ (as with tomorrow's temperature), or a more abstract space. The objects of choice are then 'acts' $a(\cdot)$: $\mathscr{S} \rightarrow \mathscr{X}$ which map states to outcomes. In the case of a finite state space, acts are usually expressed in the form $\{x_1 \text{ if } s_1;\ldots; x_n \text{ if } s_n\}$. When the state space is infinite, finite-outcome acts can be expressed in the form $a(\cdot) = [x_1 \text{ on } E_1;\ldots; x_m \text{ on } E_m]$ for some partition of $\mathscr{S}$ into a family of mutually exclusive and exhaustive 'events' $\{E_1,\ldots, E_m\}$. Except for casino games and state lotteries, virtually all real-world uncertain decisions (including all investment or insurance decisions) are made under conditions of subjective uncertainty.

In such a setting, the 'subjective expected utility hypothesis' consists of the *joint* hypothesis that the individual possesses *probabilistic beliefs*, as represented by a 'personal' or 'subjective' probability measure $\mu(\cdot)$ over the state space, and *expected utility risk preferences*, as represented by a von Neumann–Morgenstern utility function $U(\cdot)$ over outcomes, and evaluates acts according a preference function of the form $W(x_1 \text{ if } s_1;\ldots;x_n \text{ if } s_n) \equiv \sum_{i=1}^{n} U(x_i) \cdot \mu(s_i)$, $W(x_1 \text{ on } E_1;\ldots;x_m \text{ on } E_m) \equiv \sum_{i=1}^{m} U(x_i) \cdot \mu(E_i)$, or more generally, $W(a(\cdot)) \equiv \int U(a(s))\mathrm{d}\mu(s)$. Whereas all individuals facing a given *objective* prospect $\boldsymbol{P} = (x_1, p_1;\ldots; x_n, p_n)$ are assumed to 'see' the same probabilities $(p_1,\ldots, p_n)$ (though they may have different utility functions), individuals facing a given *subjective* prospect $\{x_1 \text{ if } s_1;\ldots; x_n \text{ if } s_n\}$ or $[x_1 \text{ on } E_1;\ldots; x_m \text{ on } E_m]$ will generally possess differing subjective probabilities over these states or events, reflecting their different beliefs, past experiences, and so on.

Researchers such as Arrow (1974), Debreu (1959, ch. 7) and Hirshleifer (1965, 1966) have shown how the analytics of the objective expected

utility model can be extended to both the positive and normative analysis of decisions under subjective uncertainty. As a simple example, consider an individual deciding whether to purchase earthquake insurance, and if so, how much. A simple specification of this decision involves the state space $\mathscr{S} = \{s_1, s_2\} = \{\text{earthquake; no earthquakeg}$, the individual's von Neumann–Morgenstern utility of wealth function $U(\cdot)$, their subjective probabilities $\{\mu(s_1), \mu(s_2)\}$ (which sum to unity), and the price γ of each dollar of insurance coverage. An individual with initial wealth w would then purchase q dollars' worth of coverage, where q was the solution to

$$\max_q \left[U(w - \gamma q + q) \cdot \mu(s_1) + U(w - \gamma q) \cdot \mu(s_2)\right]$$

Note that this formulation does not require that the individual and the insurance company agree on the likelihood of an earthquake.

As in the objective case, subjective expected utility can be derived from axiomatic foundations. Completeness and transitivity carry over in a straightforward way, and continuity with respect to mixture probabilities is replaced by continuity with respect to small changes in the events. The existence of additive personal probabilities is obtained by the following axiom:

Comparative likelihood: For all events A, B and outcomes $x^* \succ x$ and $y^* \succ y$, $[x^* \text{ on } A; x \text{ on } \sim A] \succcurlyeq [x^* \text{ on } B; x \text{ on } \sim B]$ implies $[y^* \text{ on } A; y \text{ on } \sim A] \succcurlyeq [y^* \text{ on } B; y \text{ on } \sim B]$.

This axiom states that if the individual 'reveals' event A to be at least as likely as event B by their preference for staking the preferred outcome x^* on A rather than on B, then this likelihood ranking will hold for all other pairs of ranked outcomes $y^* \succ y$. Finally, under subjective uncertainty the Independence Axiom is replaced by its subjective analogue, first proposed by Savage (1954):

Sure-Thing Principle: For all events E and acts $a(\cdot)$, $a^*(\cdot)$, $b(\cdot)$ and $c(\cdot)$, $[a^*(\cdot) \text{ on } E; b(\cdot) \text{ on } \sim E] \succcurlyeq [a(\cdot) \text{ on } E; b(\cdot) \text{ on } \sim E]$ implies $[a^*(\cdot) \text{ on } E; c(\cdot) \text{ on } \sim E] \succcurlyeq [a(\cdot) \text{ on } E; c(\cdot) \text{ on } \sim E]$.

where $[a(\cdot) \text{ on } E; b(\cdot) \text{ on } \sim E]$ denotes the act yielding outcome $a(s)$ for all $s \in E$ and $b(s)$ for all $s \in \sim E$.

Under subjective uncertainty, an individual's utility of outcomes might sometimes depend upon the particular state of nature. Given a health insurance decision with a state space of $\mathscr{S} = \{s_1, s_2\} = \{\text{cancer; no cancer}\}$, an individual may feel a greater need for \$100,000 in state s_1 than in state s_2. This can be modelled by means of a 'state-dependent' utility function $\{U(\cdot|\, s)|\, s \in \mathscr{S}\}$ and a'state-dependent expected utility' preference function $\hat{W}(x_1 \text{ if } s_1; \ldots; x_n \text{ if } s_n) = \sum_{i=1}^{n} U(x_i|\, s_i) \cdot \mu(s_i)$or $\hat{W}(a(\cdot)) = \int U(a(s)|\, s) d\mu(s)$. The analytics of state-dependent expected utility preferences have been extensively developed by Karni (1985).

History

The hypothesis that individuals might maximize the expectation of 'utility' rather than of monetary value was proposed independently by mathematicians Gabriel Cramer and Daniel Bernoulli, in each case as the solution to a problem posed by Daniel's cousin Nicholas Bernoulli (see Bernoulli 1738). This problem, now known as the 'St Petersburg Paradox', considers the gamble which offers a 1/2 chance of \$1, a 1/4 chance of \$2, a 1/8 chance of \$4, and so on. Although the expected value of this prospect is

$$\begin{aligned}(1/2) \cdot \$1 + (1/4) \cdot \$2 + (1/8) \cdot \$4 + \ldots\ldots \\ = \$0.50 + \$0.50 + \$0.50 + \ldots = \$\infty,\end{aligned}$$

common sense suggests that no one would be willing to forgo a very substantial certain payment in order to play it. Cramer and Bernoulli proposed that, instead of using expected value, individuals might evaluate this and other lotteries by means of their expected 'utility', with utility given by a function such as the natural logarithm or the square root of wealth, in which case the certainty equivalent of the St Petersburg gamble becomes a moderate (and plausible) amount.

Two hundred years later, the St Petersburg paradox was generalized by Karl Menger (1934), who

noted that, whenever the utility of wealth function was unbounded (as with the natural logarithm or square root functions), it would be possible to construct similar examples with infinite expected utility and hence infinite certainty equivalents (replace the payoffs \$1, \$2, \$4 ... in the above example by $x_1, x_2, x_3 \ldots$, where $U(x_i) = 2^i$ for each i). In light of this, von Neumann–Morgenstern utility functions are typically (though not universally) postulated to be bounded functions of wealth.

The earliest formal axiomatic treatment of the expected utility hypothesis was developed by Frank Ramsey (1926) as part of his theory of subjective probability, or individuals' 'degrees of belief' in the truth of alternative propositions. Starting from the premise that there exists an 'ethically neutral' proposition whose degree of belief is 1/2, and whose validity or invalidity is of no independent value, Ramsey proposed a set of axioms on how the individual would be willing to stake prizes on its truth or falsity, in a manner which allowed for the derivation of the 'utilities' of these prizes. He then used these utility values and betting preferences to determine the individual's degrees of belief in other propositions. Perhaps because it was intended as a contribution to the philosophy of belief rather than to the theory of risk bearing, Ramsey's analysis did not have the impact among economists that it deserved.

The first axiomatization of the expected utility model to receive widespread attention was that of John von Neumann and Oskar Morgenstern, presented in connection with their formulation of the theory of games (von Neumann and Morgenstern 1944, 1947, 1953). Although this development was recognized as a breakthrough, the mistaken belief that von Neumann and Morgenstern had somehow mathematically overthrown the Hicks–Allen 'ordinal revolution' led to some confusion until the difference between 'utility' in the von Neumann-Morgenstern and the ordinal (that is, non-stochastic) senses was illuminated by writers such as Ellsberg (1954) and Baumol (1958).

Another factor which delayed the acceptance of the theory was the lack of recognition of the role played by the Independence Axiom, which did not explicitly appear in the von Neumann–Morgenstern formulation. In fact, the initial reaction of researchers such as Baumol (1951) and Samuelson (1950) was that there was no reason why preferences over probability distributions must *necessarily* be linear in the probabilities. However, the independent discovery of the Independence Axiom by Marschak (1950), Samuelson (1952) and others, and Malinvaud's (1952) observation that it had been implicitly invoked by von Neumann and Morgenstern, led to an almost universal acceptance of the expected utility hypothesis as both a normative and positive theory of behaviour towards risk. This period also saw the development of the elegant axiomatization of Herstein and Milnor (1953) as well as Savage's (1954) joint axiomatization of utility and subjective probability, which formed the basis of the state-preference approach described above.

While the 1950s essentially saw the completion of foundational work on the expected utility model, subsequent decades saw the flowering of its analytic capabilities and its application to fields such as portfolio selection (Merton 1969), optimal savings (Levhari and Srinivasan 1969; Fleming and Sheu 1999), international trade (Batra 1975; Lusztig and James 2006), environmental economics (Wolfson et al. 1996), medical decision-making (Meltzer 2001) and even the measurement of inequality (Atkinson 1970). This movement was spearheaded by the development of the Arrow–Pratt characterization of risk aversion (see above) and the characterization, by Rothschild–Stiglitz (1970, 1971) and others, of the notion of 'increasing risk'. This latter work in turn led to the development of a general theory of 'stochastic dominance' (for example, Whitmore and Findlay 1978; Levy 1992), which has further expanded the analytical powers of the model.

Although the expected utility model received a small amount of experimental testing by economists in the early 1950s (for example, Mosteller and Nogee 1951; Allais 1953) and continued to be examined by psychologists, economists' interest in the empirical validity of the model waned from the mid-1950s through the mid-1970s, no doubt due to both the normative appeal of the Independence Axiom and model's analytical successes.

However, since the late 1970s there has been a revival of interest in the testing of the expected utility model; a growing body of evidence that individuals' preferences *systematically* depart from linearity in the probabilities; and the development, analysis and application of alternative models of choice under objective and subjective uncertainty. It is fair to say that today the debate over the descriptive (and even normative) validity of the expected utility hypothesis is more extensive than it has been in over half a century, and the outcome of this debate will have important implications for the direction of research in the economics of uncertainty.

See Also

- ▶ Bernoulli, Daniel (1700–1782)
- ▶ Non-Expected Utility Theory
- ▶ Ramsey, Frank Plumpton (1903–1930)
- ▶ Risk
- ▶ Risk Aversion
- ▶ Savage's Subjective Expected Utility Model
- ▶ Uncertainty
- ▶ Utility

Bibliography

Allais, M. 1953. Fondements d'une théorie positive des choix comportant un risque et critique des postulats et axiomes de l'école Américaine. *Colloques Internationaux du Centre National de la Recherche Scientifique* 40: 257–332. Trans. as: The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School. In *Expected utility hypotheses and the Allais paradox*, ed. M. Allais and O. Hagen. Dordrecht: D. Reidel, 1979.

Arrow, K. 1974. *Essays in the theory of risk-bearing*. Amsterdam: North-Holland.

Atkinson, A. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Batra, R. 1975. *The pure theory of international trade under uncertainty*. London: Macmillan.

Baumol, W. 1951. The Neumann–Morgenstern utility index: An ordinalist view. *Journal of Political Economy* 59: 61–66.

Baumol, W. 1958. The cardinal utility which is ordinal. *Economic Journal* 68: 665–672.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*. Trans. as Exposition of a new theory on the measurement of risk. *Econometrica* 22 (1954): 23–36.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New Haven: Yale University Press.

Ellsberg, D. 1954. Classical and current notions of 'measurable utility'. *Economic Journal* 64: 528–556.

Fishburn, P. 1982. *The foundations of expected utility*. Dordrecht: D. Reidel.

Fleming, W., and S.-J. Sheu. 1999. Optimal long term growth rate of expected utility of wealth. *Annals of Applied Probability* 9: 871–903.

Friedman, M., and L. Savage. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56: 279–304.

Herstein, I., and J. Milnor. 1953. An axiomatic approach to measurable utility. *Econometrica* 21: 291–297.

Hey, J. 1979. *Uncertainty in microeconomics*. Oxford/New York: Martin Robinson/New York University Press.

Hirshleifer, J. 1965. Investment decision under uncertainty: Choice theoretic approaches. *Quarterly Journal of Economics* 79: 509–536.

Hirshleifer, J. 1966. Investment decision under uncertainty: Applications of the state-preference approach. *Quarterly Journal of Economics* 80: 252–277.

Hirshleifer, J., and J. Riley. 1979. The analytics of uncertainty and information – An expository survey. *Journal of Economic Literature* 17: 1375–1421.

Karni, E. 1985. *Decision making under uncertainty: The case of state-dependent preferences*. Cambridge, MA: Harvard University Press.

Karni, E., and D. Schmeidler. 1991. Utility theory with uncertainty. In *Handbook of mathematical economics*, ed. W. Hildenbrand and H. Sonnenschein, vol. 4. Amsterdam: North-Holland.

Kreps, D. 1988. *Notes on the theory of choice*. Boulder: Westview Press.

Levhari, D., and T.N. Srinivasan. 1969. Optimal savings under uncertainty. *Review of Economic Studies* 36: 153–164.

Levy, H. 1992. Stochastic dominance and expected utility: Survey and analysis. *Management Science* 38: 555–593.

Lippman, S., and J. McCall. 1981. The economics of uncertainty: Selected topics and probabilistic methods. In *Handbook of mathematical economics*, ed. K. Arrow and M. Intriligator, vol. 1. Amsterdam: North-Holland.

Lusztig, M., and P. James. 2006. How does free trade become institutionalised? An expected utility model of the Chrétien era. *World Economy* 29: 491–505.

Machina, M. 1983. *The economic theory of individual behavior toward risk: Theory, evidence and new directions*. Technical report no. 433. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Malinvaud, E. 1952. Note on von Neumann–Morgenstern's strong independence axiom. *Econometrica* 20: 679–680.

Markowitz, H. 1952. The utility of wealth. *Journal of Political Economy* 60: 151–158.
Marschak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18: 111–141.
Meltzer, D. 2001. Addressing uncertainty in medical cost-effectiveness analysis: Implications of expected utility maximization for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research. *Journal of Health Economics* 20: 109–129.
Menger, K. 1934. Das Unsicherheitsmoment in der Wertlehre. *Zeitschrift für Nationalökonomie*. Trans. as: The role of uncertainty in economics. In *Essays in mathematical economics in honor of Oskar Morgenstern*, ed. M. Shubik. Princeton: Princeton University Press, 1967.
Merton, R. 1969. Lifetime portfolio selection under uncertainty: The continuous time case. *Review of Economics and Statistics* 51: 247–257.
Mosteller, F., and P. Nogee. 1951. An experimental measurement of utility. *Journal of Political Economy* 59: 371–404.
Pratt, J. 1964. Risk aversion in the small and in the large. *Econometrica* 32: 122–136.
Quirk, J., and R. Saposnick. 1962. Admissibility and measurable utility functions. *Review of Economic Studies* 29: 140–146.
Ramsey, F. 1926. Truth and probability. In *The foundations of mathematics and other logical essays*, ed. R. Braithwaite. New York: Harcourt, Brace and Co, 1931. Reprinted in *Foundations: Essays in philosophy, logic, mathematics and economics*, ed. D. Mellor. New Jersey: Humanities Press, 1978.
Ross, S. 1981. Some stronger measures of risk aversion in the small and in the large, with applications. *Econometrica* 49: 621–638.
Rothschild, M., and J. Stiglitz. 1970. Increasing risk I: A definition. *Journal of Economic Theory* 2: 225–243.
Rothschild, M., and J. Stiglitz. 1971. Increasing risk II: Its economic consequences. *Journal of Economic Theory* 3: 66–84.
Samuelson, P. 1950. Probability and attempts to measure utility. *Economic Review* 1: 167–173.
Samuelson, P. 1952. Probability, utility, and the independence axiom. *Econometrica* 20: 670–678.
Savage, L. 1954. *The foundations of statistics*. New York: John Wiley & Sons. Revised edition: New York: Dover, 1972.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*. 2nd ed. Princeton: Princeton University Press.
von Neumann, J., and O. Morgenstern. 1953. *Theory of games and economic behavior*. 3rd ed. Princeton: Princeton University Press.
Whitmore, G., and M. Findlay, eds. 1978. *Stochastic dominance: An approach to decision making under risk*. Lexington: D.C. Heath.
Wolfson, L., J. Kadane, and M. Small. 1996. Expected utility as a policy making tool: An environmental health example. In *Bayesian biostatistics*, ed. D. Berry and D. Stangl. New York: Marcel Dekker.

Expenditure Tax

A. P. Thirlwall

The idea of an expenditure tax has a long ancestry, dating back at least to Hobbes, who argued that people should be taxed according to the resources of the community they absorb not according to what they contribute. The case was later taken up by J.S. Mill, Marshall, Pigou and Irving Fisher. In modern times, the advocacy of an expenditure tax is most associated with the Cambridge economist Nicholas Kaldor (1955). Recently it has been espoused by the Meade Committee (Meade 1978), and separately by two members of that Committee, Kay and King (1978).

There are efficiency and equity arguments for considering an expenditure tax as an alternative to income taxation. As far as efficiency is concerned, there is a commonly held view that because income tax involves the double taxation of saving, and therefore lowers the rate of return on saving below the rate of return on investment, this distorts the choice between consumption and saving, and represents a wasteful distortion. However, recent work, using optimal tax theory, has questioned this conclusion. The theory of the second best teaches that alternative tax systems cannot be evaluated according to the number of distortions: the magnitude of the distortions and their interaction also needs to be taken into account. Using optimal tax theory in an intertemporal context, Atkinson and Sandmo (1980) have shown that no firm conclusions can be reached on the relative efficiency of expenditure and income taxes from a welfare point of view: it all depends on the form of the social

welfare function, what other instruments governments can use to achieve a desired intertemporal allocation of consumption, and on crucial parameters of the model such as the interest elasticity of labour supply.

As far as equity is concerned, the case for an expenditure tax may be stronger. The principle of progressive income taxation rests on the concept of taxable capacity or ability to pay. The question is, does 'income' approximate to this concept? There are three main difficulties. First, income is only one measure of taxable capacity. Secondly, income itself is not an unambiguous concept. Thirdly, the actual definition of income for tax purposes can introduce inequities into the system by some receipts being treated as income and others not. Income is taken as a proxy for 'spending power', but there are other sources of spending power (e.g. wealth) and it is not easy to express them all in a single measure of taxable capacity. There are particular problems associated with irregular receipts and capital gains. It can be argued that many of the problems created by the non-comparability of different forms of income, wealth and capital gains would be resolved by taxing expenditure rather than income. The individual himself would declare his spending power when he spends. Since there is no objective definition of income that can provide a true measure of spending power, there can be no presumption that any income tax system would be superior from an equity point of view to an expenditure tax.

Kaldor was the first to argue in a comprehensive way that the measurement of income as a measure of taxable capacity is inevitably ambiguous and is likely to be a bad proxy for the measurement of spending power, so that the taxation of spending as such may be regarded at least as equitable as income tax, if not more so, with other positive advantages – particularly, an expenditure tax would be a more efficient instrument for controlling the economy, so that there is no necessary conflict between an egalitarian system of taxation, efficiency and growth. The Meade Committee, as well as mentioning the traditional arguments concerning the difficulty of defining income and measuring accruals, placed most emphasis on the elimination of capital market distortions, particularly those associated with various concessions in the existing income tax system which have differential and distorting effects on rates of return to saving outlets, and with having to correct nominal capital gains and losses for inflation. Such problems automatically disappear with an expenditure tax.

Kaldor also discussed at length the effect that a switch to an expenditure tax is likely to have on risk bearing, the supply of effort, saving and economic progress. It turns out that it is impossible to predict the net effects with any degree of certainty. As far as risk bearing is concerned it is impossible to say whether an expenditure tax is better or worse than an income tax yielding the same revenue, since on the one hand it is less discriminating against risk in so far as part of taxable income is saved, but on the other hand is more discriminating in so far as part of the capital gain is spent. Likewise, as far as the supply of effort is concerned, it is possible to reach different conclusions according to the assumptions made concerning the relative stability of income and consumption, and whether taxation is progressive or proportional. Kaldor did not pay much attention to the argument that an expenditure tax would avoid the double taxation of saving under income tax, and therefore avoid distortions and encourage saving, but it has other advantages relating to enterprise and economic progress. For example, without a capital gains tax, income tax puts a premium on speculation compared with an expenditure tax where both yield and capital gains are equally taxed if spent, or equally exempt if saved. An expenditure tax which discouraged speculation would enhance the supply of risk capital.

Several theoretical objections have been raised against the expenditure tax, but none is very convincing. The main difficulty concerns the practical implementation of the tax. In evidence to the 1929 Colwyn Committee on National Debt and Taxation, Keynes had earlier described the expenditure tax idea as theoretically sound but 'practically impossible'. This was the prevailing view (see also Pigou 1928) largely because of the difficulty of getting taxpayers to keep accurate records of personal expenditure and checking returns. It was Irving Fisher (1937) who first showed that this

would not be necessary since a person's expenditure is the difference between what he has available for spending and what he has left at the end of the accounting period. Thus in theory the only information required is the size of a person's bank balance at the beginning of the year plus income and other receipts, and from this is then deducted net investments, exempted expenditure and the size of the bank balance at the end of the year, and the difference is chargeable expenditure. The major problems concern the definition of chargeable expenditure, and evasion through the avoidance of the use of bank accounts.

In his original exposition of the expenditure tax, Kaldor perhaps overestimated the drawbacks of income tax and understated the difficulties of the expenditure tax. Conceptually the dividing line between what is consumption and what is saving may be said to be as arbitrary and fraught with difficulties, as in answering the question, when does income accrue? In defence, he admitted, however, that comparing the expenditure tax with a more *comprehensive* income tax, the balance between the two is much more finely poised: the conclusion which Prest (1979) also comes to in his review of the Meade Committee Report.

See Also

- ▶ Consumption Taxation
- ▶ Direct Taxes
- ▶ Public Finance
- ▶ Taxation of Income
- ▶ Taxation of Wealth

Bibliography

Atkinson, A.B.., and A. Sandmo. 1980. Welfare implications of the taxation of savings. *Economic Journal* 90: 529–549.

Fisher, I. 1937. Income in theory and income taxation in practice. *Econometrica* 5: 1–55.

Kaldor, N. 1955. *An expenditure tax*. London: Allen and Unwin.

Kay, J.A., and M.A. King. 1978. *The British tax system*. London: Oxford University Press.

Meade, J. 1978. *The structure and reform of direct taxation*. London: George Allen and Unwin.

Pigou, A.C. 1928. *A study in public finance*. London: Macmillan.

Prest, A.R. 1979. The structure and reform of direct taxation. *Economic Journal* 89: 243–260.

Experimental Economics

Vernon L. Smith

E

Abstract

Experimental methods have features in common across all the sciences. All tend to use the framework of falsification, but there is inherent ambiguity in knowing which of the many hypotheses necessary to construct a test are negated by observations contrary to predictions. This ambiguity tends to engender much discussion, contestability and the design of new experiments that attempt to resolve the open qsts. This social process is not part of the logic of scientific testing, but it explains what scientists do and how new results become established.

Keywords

Bounded rationality; Brownian motion; Constructivist rationality; Dictator games; Double auction; Duhem–Quine problem; Ecological rationality; Equilibrium; Excess demand; Experimental economics; Experimental knowledge; Extensive form games; Falsificationism; First-price auctions; Friedman, M.; Hayek, F. A.; Internet, economics of; Market power; Mechanism design; Neuroscience; Perfect competition; Perfect information; Positive economics; Prediction; Private information; Repeated games; Risk neutrality; Strategic games; Subjective probability; Subjective utility; Testing; Ultimatum games

JEL Classifications

C9

> But I believe that there is no philosophical highroad in science, with epistemological signposts. . .we are in a jungle and find our way out by trial and error, building our road *behind* us as we proceed. We do not *find* signposts at crossroads, but our scouts *erect them,* to help the rest.
> —Max Born, *Experiment and Theory in Physics* (1943)

> . . . they were criticized [those studying observational learning in a social context] for being unscientific and performing *uncontrolled* experiments. In science, there's nothing 'worse' than an experiment that's uncontrolled.
> —Temple Grandin, *Animals in Translation* (2005, bracketed comments added).

The subject matter of this article is rationality in science particularly as it applies to experimental methods. In this context 'rationality' is commonly used to refer to a particular conception that Hayek (1967, p 85) has called:

> *Constructivist Rationality,* which, applied to individuals, associations or organizations, involves the conscious deliberate use of reason to analyze and prescribe actions judged to be better than alternative feasible actions that might be chosen; applied to institutions it involves the deliberate design of rule systems to achieve desirable performance. The latter include 'optimal design' where the intention is to provide incentives for agents to choose better actions than would result from alternative arrangements.

Rationality in socioeconomic systems, including scientific communities, cannot be adequately understood by restricting one's perspective to this traditional Cartesian framework. In the discussion that follows I want to draw upon a second conception of rationality:

> *Ecological rationality* refers to emergent order in the form of practices, norms and rules governing action by individuals, groups and institutions that are part of our cultural and biological heritage, created by human interactions, but not by conscious human design.

I have argued (Smith 2003) that rationality in the economy depends on individuals whose behaviour is conditioned by cultural norms and emergent institutions that evolve from human experience, neither of which is ultimately derived from constructivist reason; although, clearly, constructivist ideas are important sources of variation for the gristmill of ecological selection. Parallel considerations apply to rationality in scientific method, the extension to be treated here.

Stated briefly, here is the argument that I will present: scientific methodology reveals a predominately constructivist theme largely guided by the following:

- falsification criteria for hypotheses derived from theories;
- experimental designs for testing hypotheses;
- statistical tests; and
- liturgies of reporting style that have become standard in scientific papers.

But all tests of theory are necessarily joint tests of hypotheses derived from theory and the set of auxiliary hypotheses necessary to implement, construct and execute the tests: this is the well-known Duhem–Quine (D–Q) problem. Thus, whatever might be the testing rhetoric of scientists, they do not reject hypotheses, and their antecedent theories, on the basis of falsifying outcomes. But this is not cause for despair, let alone retreat into a narrow postmodern sea of denial in which science borders on unintended fraud. D–Q is a property of inquiry, a truth, and as such a source for deepening our understanding of what *is,* not a clever *touché* for exposing the rhetorical pretensions of science. But the failure of all philosophy of science programs to articulate a rational constructivist methodology of science that serves to guide scientists, or explain what they do, as well as what they say about what they do, does not mean that science is devoid of rationality or that scientific communities fail to generate rational programmes of scientific inquiry. Thus, scientists engage in commentary, reply, rebuttal and vigorous discussions over whether the design is appropriate, the test adequate, whether the procedures and measurements might be flawed, and the conclusions and inpts correct. One must look to this conversation in the scientific community in asking whether and how science sorts out competing primary and auxiliary hypotheses after each new set of tests results are made available. If this conversation does not read like a theorem and its proof, and fails to reduce methodology to a consciously rigorous science of

inquiry, this is because we can never reduce the testing enterprise to a simple up or down test of an isolated non-trivial hypothesis; so be it.

If emergent method is rational in science it must be a form of ecological rationality; this means that it rightly and inevitably grows out of the norms, practices and conversation that characterize meaningful interactions in the scientific community. Listen not only to what scientists say about what they do, ignoring the arrogant tone in their little knowledge, but also examine what they do. The power behind the throne of accomplishment in the human career is our *sociality*, and the unintended mansions that are built by that sociality. The long view of that career is in sharp focus: our accumulation of knowledge and its expression in technology enabled us to survive the Pleistocene, people the Earth, penetrate the heavens, and explore the ultimate particles and forces of matter, energy and life. That achievement hardly deserves to be described as either irrational or non-rational.

What does it mean to test a theory or a hypothesis derived from a theory? Scientists everywhere say and believe that the unique feature of science is that theories, if they are to be acceptable, require rigorous support from facts based on replicable observations. But the deeper one examines this belief the more elusive becomes the task of giving it precise meaning and content in the context of conventional rational programs of inquiry.

Can We Derive Theory Directly from Observations?

Prominent scientists through history have believed that the answer to this question is 'yes', and that this was their modus operandi. Thus, quoting from two of the most influential scientists in history:

> I frame no hypotheses; for whatever is not deduced from the phenomena … have no place in experimental philosophy … (in which) … particular propositions are inferred from the phenomena … (Isaac Newton, *Principia,* 1687; quoted in Segrè 1984, p. 66)

> Nobody who has really gone into the matter will deny that in practice the world of phenomena uniquely determines the theoretical system, in spite of the fact that there is no theoretical bridge between phenomena and their theoretical principles. (Einstein 1934, pp. 22–3)

But these statements are without validity. 'One can today easily demonstrate that there can be no valid derivation of a law of nature from any finite number of facts' (Lakatos 1978, vol. 1, p. 2). Yet in economics critics of standard theory call for more empirical work on which to base development of *the* theory, and generally the idea persists that the essence of science is its rigorous observational foundation.

But how are the facts and theories of science to be connected so that each constructively informs and enriches the other?

Newton passionately believed not just that he was proffering lowly hypotheses, but that his laws were derived directly, by logic, from Kepler's discovery that the planets moved in ellipses. But Newton only showed that the path was an ellipse if there are $n = 2$ planets. Kepler was wrong in thinking that the planets followed elliptical paths, and to this day there is no solution for the $n(>2)$-body problem, and in fact the paths can be chaotic. Thus, when he published the *Principia,* Newton's model could not account for the motion of our nearest and most accurately observable neighbour, the moon, whose orbit is strongly influenced by both the sun and the earth.

Newton's sense of his scientific procedure is commonplace: one studies an empirical regularity (for example, the 'trade-off' between the rate of inflation and the unemployment rate), and proceeds to articulate a model from which a functional form can be derived that yields the regularity. In the above confusing quotation, Einstein seems to agree with Newton. At other times he appears to articulate the more qualified view that theories make predictions, which are then to be tested by observations (see his insightful comment below on Kaufmann's test of special relativity theory), while on other occasions his view is that reported facts are irrelevant compared to theories based on logically complete meta theoretical principles, coherent across a broad spectrum of

fundamentals (see Northrup 1969, pp. 387–408). Thus, upon receiving the telegraphed news that Eddington's 1919 eclipse experiments had 'confirmed' the general theory, Einstein showed it to a doctoral student who was jubilant, but he commented unmoved: 'I knew all the time that the theory was correct.' But what if it had been refuted? 'In that case I'd have to feel sorry for God, because the theory is correct' (quoted in Fölsing 1997, p. 439).

The main theme I want to develop in this and subsequent sections is captured by the following quotation from a lowbrow source, the mythical character Phaedrus in *Zen and the Art of Motorcycle Maintenance,* '. . . the number of rational hypotheses that can explain any given phenomena is infinite' (Pirsig 1981, p. 100).

Proposition 1 Particular hypotheses derived from any testable theory imply certain observational outcomes; the converse is false (Lakatos 1978, vol. 1, pp. 2, 16, *passim*).

Theories produce mathematical theorems. Each theorem is a mapping from postulated statements (assumptions) into derived or concluding statements (the theoretical results). Conventionally, the concluding statements are what the experimentalist uses to formulate specific hypotheses (models) that motivate the experimental design that is implemented. The conditions that underpin the hypotheses are the objects of control in an economics experiment, insofar as they can be controlled. Since not every assumption can always be reproduced in the experimental design the problem of the 'controlled experiment' is one of trying to minimize the risk that the results will fail to be interpretable as a test of the theory because one or more assumptions were violated. An uncontrolled assumption that is postulated to hold in interpreting test results is one of many possible contingent auxiliary hypotheses to be discussed below.

The wellspring of testable hypotheses in economics and game theory is to be found in the marginal conditions defining equilibrium points or strategy decision functions that constitute a theoretical equilibrium. In games against nature the subject-agent is assumed to choose among alternatives in the feasible set that which maximizes his or her outcome (reward, utility or payoff), subject to the technological and other constraints on choice. Strategic games are solved by the device of reducing them to games against nature, as in a non-cooperative (Cournot–Nash) equilibrium (pure or mixed) where each agent is assumed to maximize his or her own outcome, given (subject to the constraints of) the maximizing behaviour of all other agents. The equilibrium strategy when used by all but agent i reduces i's problem to an own maximizing choice of that strategy. Hence, in economics, all testable hypotheses come from the marginal conditions (or their discrete opportunity cost equivalent) for maximization that define equilibrium for an individual or across individuals interacting through an institution. These conditions are implied by the theory from which they are derived, but given experimental observations consistent with (that is, supporting) these conditions there is no way to reverse the steps used to derive the conditions, and deduce the theory from a set of observations on subject choice. Behavioural point observations conforming to an equilibrium theory cannot be used to deduce or infer either the equations defining the equilibrium or the logic and assumptions of the theory used to derive the equilibrium conditions.

Suppose, however, that the theory is used to derive non-cooperative best reply functions for each agent that maps one or more characteristics of each individual into that person's equilibrium decision. Suppose next that we perform many repetitions of an experiment varying some controllable characteristic of the individuals, such as their assigned values for an auctioned item, and obtain an observed response for each value of the characteristic. This repetition of course must be assumed always to support equilibrium outcomes. Finally, suppose we use this data to estimate response functions obtained from the original maximization theory. First order conditions defining an optimum can always be treated formally as differential equations in the original criterion function. Can we solve these equations and 'deduce' the original theory?

An example is discussed in Smith et al. (1991) from first-price auction theory. Briefly the idea is this. Each of N individuals in a repeated auction

game is assigned value $v_i(t)$ ($i = 1,\ldots, N;\quad t = 1,2,\ldots,T$) from a distribution, and on each trial, t, i bids $b_i(t)$. On each trial each i is assumed to choose a bid that maximizes expected utility.

$$\max_{0\le b_i\le v_i} (v_i - b_i)^{r_i} G_i(b_i) \tag{1}$$

where r_i ($0 < r_i \le 1$) is i's measure of constant relative risk aversion, and $G_i(b_i)$ is i's probability belief that a bid of b_i will win. This leads to the first order condition,

$$(v_i - b_i)G_i'(b_i) - r_i G_i(b_i) = 0. \tag{2}$$

If all i have identical common rational probability expectations

$$G_i(b_i) = G(b_i). \tag{3}$$

This leads to a closed form equilibrium bid function (see Cox et al. 1988).

$$\begin{aligned} b_i &= (N-1)v_i/(N-1+r_i), \quad b_i \leqq b \\ &= 1 - G(b)/G'(b), \quad \text{for all } i. \end{aligned} \tag{4}$$

The data from experimental auctions strongly support linear subject bid rules of the form

$$b_i = \alpha_i v_i, \tag{5}$$

obtained by linear regression of b_i on v_i using the T observations on (b_i, v_i), for given N, with $\alpha_i = (N - 1)v_i/(N - 1 + r_i)$. Can we reverse the above steps, then integrate (5) to get (1)? The answer is 'no'. Equation (2) can be derived from maximizing either (1) or the criterion

$$(v_i - b_i)G(b_i)^{1/r}, \tag{1'}$$

in which subjective probabilities, rather than profit, are 'discounted' in computing the expectation. That is, without all the assumptions used to get (4), we cannot uniquely conclude (1). In (1′) we have $G_i(b_i) = G(b_i)^{1/r_i}$ instead of (3), and this is not ruled out by the data.

In fact, instead of (1) or (1′) we could have maximized

$$(v_i - b_i)^{\beta_i} G(b_i)^{\beta_i/r_i}, \quad \text{with} \quad r_i \le \beta_i \le 1 \tag{1'}$$

giving an infinite mixture of subjective utility and subjective probability models of bidding.

There is a special case of the above model that is reversible: all bidders are risk neutral with, $r_i = 1$. While that model is often defended in the abstract, and is the workhorse assumption for deriving theorems under uncertainty, it fails all the lab and field empirical tests known to me. Risk neutrality trivializes decisions by requiring all humans to have identical preferences. It fails because people are inherently heterogeneous in making decisions under uncertainty, and this empirical diversity is captured by an appearance (or a mirage) in the data of non-neutral 'risk'. Equation (1′) above provides a hint of the manner in which individual diversity can appear in data inpts that confound measures of risk with other sources of heterogeneity.

Thus, in general, we cannot backward infer from empirical equilibrium conditions, even when we have a large number of experimental observations, to arrive at the original parameterized model within the general theory. *The purpose of theory is precisely one of imposing much more structure on the problem than can be inferred from the data.* This is because the assumptions used to deduce the theoretical model contain more information, such as (3), than the data – the theory is underdetermined.

Economics: Is It An Experimental Science?

All editions of Paul Samuelson's *Principles of Economics* refer to the inability of economists to perform experiments. This continued for a short time after William Nordhaus joined Samuelson as a coauthor. Thus, 'Economists . . . cannot perform the controlled experiments of chemists and biologists because they cannot easily control other important factors' (Samuelson and Nordhaus 1985, p. 8).

My favourite quotation, however, is supplied by one of the 20th century's foremost Marxian

economists, Joan Robinson. To wit, 'Economists cannot make use of controlled experiments to settle their differences' (Robinson 1979, p. 1319). Like Samuelson, she was not accurate – economists do indeed perform controlled experiments – but how often have they, or their counterparts in any science, used them to 'settle their differences?' Here she was expressing the popular image of science, which is indeed one in which 'objective' facts are the arbiters of truth that in turn 'settle' differences. The caricatured image is that of two scientists, who, disagreeing on a fundamental principle, go to the lab, do a 'crucial experiment', and learn which view is assuredly right. The hypothesis they are testing is not underdetermined by the test data. They do not argue about the result; their question is answered and they move on to a new topic that is not yet 'settled.'

Although these quotations provide telling commentaries on the state of the profession's knowledge of the development of experimental methods in economics since the 1950s, there is a deeper question of whether there are more than a very small number of non-experimentalists in economics that understand key features of our methodology. These are twofold: (*a*) employ a reward scheme to motivate individual behaviour in the laboratory within an economic environment defining gains from trade that are controlled by the experimenter – for example, the supply of and demand for an abstract item in an isolated market or an auction; and (*b*) use the observations to test predictive hypotheses derived from one or more models (formal or informal) of behaviour in these environments using the rules of a particular trading institution – for example, the equilibrium clearing price and corresponding exchange volume when subjects trade under some version of an oral or electronic double auction, posted pricing, sealed bidding, and so on. This differs from the way that economics is commonly researched, taught and practised, which implies that it is largely an a priori science in which economic problems come to be understood by thinking about them. This generates logically correct, internally consistent theories and models. The data of econometrics are then used for 'testing' between alternative model specifications within basic equilibrium theories that are not subject to challenge, or to estimate the supply and/or demand parameters assumed to generate data representing equilibrium outcomes by an unspecified process. (Leamer 1978, and others have challenged the interpretation of this standard econometric methodology as a scientific 'testing' programme as distinct from a programme for specification searches of data.) Theories are not so much subject to doubt as used to impose restrictions on the data that allow parameters to be estimated. *Its constructivism all the way down.*

I want to report two examples indicating how counter-intuitive it has been for prominent economists to see the function of laboratory experiments in economics. The first example is contained in a quotation from Hayek whose Nobel citation was for his theoretical conception of the price system as an information system for coordinating agents with dispersed information in a world where no single mind or control centre possesses, or can ever have knowledge of, this information. His critique and rejection of mainstream quantitative methods, 'scientism', in economics are well known (see, for example, Hayek 1942, 1945). But in his brilliant paper interpreting competition as a discovery process, rather than a model of equilibrium price determination, he argues:

> . . . wherever the use of competition can be rationally justified, it is on the ground that we do not know in advance the facts that determine the actions of competitors . . . competition is valuable only because, and so far as, its results are unpredictable and on the whole different from those which anyone has, or could have, deliberately aimed at. . . . The necessary consequence of the reason why we use competition is that, in those cases in which it is interesting, the validity of the theory can never be tested empirically. We can test it on conceptual models, and *we might conceivably test it in artificially created real situations, where the facts that competition is intended to discover are already known to the observer. But in such cases it is of no practical value, so that to carry out the experiment would hardly be worth the expense.* (F.A. Hayek 1978, p. 255; emphasis added)

Hayek describes with clarity an important use (unknown to him) that has been made of experiments – testing competitive theory 'in

artificially created real situations, where the facts which competition is intended to discover are already known to the observer' – then proceeds to completely fail to see how such an experiment could be used to test his own proposition that competition is a discovery procedure, under the condition that neither agents as a whole nor any one mind needs to know what each agent knows. Rather, his concern for dramatizing what is arguably the most important socio-economic idea of the 20th century seems to have caused him to interpret his suggested hypothetical experiment as 'of no practical value' since it would (if successful) merely reveal what the observer already knew!

I find it astounding that one of the most profound thinkers in the 20th century did not see the demonstration potential and testing power of the experiment he suggests for testing the proposition: with competition *no one in the market* need know in advance the actions of competitors, and that competition is valuable only because, and so far as, its results are unpredictable *by anyone in the market* and on the whole different from those which anyone *in the market* has, or could have, deliberately aimed at. Yet, unknown to me at the time, this is precisely what my first experiment conducted in January 1956, published later as 'Test 1', was all about (Smith 1962).

I assembled a considerable number of experiments for a paper 'Markets as Economizers of Information: Experimental Examination of the "Hayek Hypothesis"', presented at the 50th Jubilee Congress of the Australian and New Zealand Association for the Advancement of Science, in Adelaide, Australia, 12–16 May 1980. A version of this paper was reprinted in Smith (1991, pp. 221–35). Here is what I called the Hayek Hypothesis. Strict privacy together with the trading rules of a market institution (the oral double auction in this case) is sufficient to produce efficient competitive market outcomes. The alternative was called the Complete Knowledge Hypothesis: competitive outcomes require perfectly foreseen conditions of supply and demand, a statement attributable to many economists, including Paul Samuelson who refers to 'foreseen changes in supply and demand' (Samuelson 1966, p. 947 and *passim),* that can be traced back to W.S. Jevons in 1871. (Stigler 1957, provides a historical treatment of the concept of perfect competition.) In this empirical comparison the Hayek Hypothesis was strongly supported. This theme had been visited earlier (before I had become aware or at least fully appreciative of Hayek's 1945 contribution that equilibrium theory was a tautology) in Smith (1976), wherein eight experiments comparing private information with complete information showed that complete information was neither necessary nor sufficient for convergence to a competitive equilibrium: complete information interfered with, and slowed, convergence compared with private information. Shubik (1959, pp. 169–71) had noted earlier, and correctly, the confusion inherent in ad hoc claims that perfect knowledge is a requirement of pure (or sometimes perfect) competition. The experimental proposition that private information increases support for noncooperative, including competitive, outcomes applies not only to markets but also to the two-person extensive-form repeated games reported by McCabe et al. (1998). Hence it is clear that without knowledge of the other's payoff it is not possible for players to identify and consciously coordinate on a cooperative outcome. Thus, as we have learned, payoff information is essential to conscious coordination in two-person interactions, but irrelevant, if not pernicious, in impersonal market exchange. We note in passing that the large number of experiments demonstrating the Hayek Hypothesis in no sense implies that there may not be exceptions (Holt 1989). It's the other way around: this large set of experiments demonstrates clearly that there are exceptions almost everywhere to the Complete Knowledge Hypothesis, and these exceptions were not part of a prior design created to provide a counter example.

Holt and his coauthors have asked, 'Are there any conditions under which double-auction markets do not generate competitive outcomes? The only exception seems to be an experiment with a "market power" design reported by Holt et al. (1986) and replicated by Davis and Williams (1991)' (Davis and Holt 1993, p. 154; also see Holt 1989). The example reported in this exception was a market in which there was a constant

excess supply of only one unit – a market with inherently weak equilibrating properties. Actually, there were two earlier reported exceptions, neither of which required market power: (*a*) one in which information about private circumstances is known by all traders (the alternative to the Hayek Hypothesis as stated above), and (*b*) an example in which the excess supply in the market was only two units. Exception (*a*) was reported in Smith (1980, 1991, pp. 104–5) and (*b*) in Smith (1991, p. 67). The above cited exceptions, attributed to market power, would need to be supplemented with comparisons in which there was just one unit of excess demand, but no market power, in order to show whether or not each exception was driven by market power, and not the fact that there is only one unit of excess supply which may be enabling of above equilibrium prices even is there is no market power. Sometimes missing in the standard toolkit of experimentalists are routines for challenging our own interpretation of data where there are confounding elements in the explanations of the results. But in this respect our methodology keeps getting better.

My second example involves the same principle as the first. It derives from a personal conversation in the early 1980s with one of my favorite Nobel Laureates in economics, a prominent theorist. In response to a qst, I described the experimental public goods research I had been doing in the late 1970s and early 1980s comparing the efficacy of various public good mechanisms: Lindahl, Groves–Ledyard, the so- called auction election or mechanism. (See the public goods papers reprinted in Smith 1991.) He wondered how I had achieved control over the efficient allocation as the benchmark used in these comparisons. So I explained what I had naively thought was commonly understood by then: I give each subject a payoff function (table) in monetary payoffs defined jointly over variable units of a public (common outcome) good, and variable units of a private good. This allows the experimenter to solve for the social optimum and then use the experimental data to judge the comparative performance of alternative public good incentive mechanisms. Incredibly, he objected that if, as the experimenter, I have sufficient information to know what constitutes the socially optimal allocation then I did not need a mechanism! I can just impose the optimal allocation! Baldly stated, economics is about deducing best actions from theory, not finding ways to test its propositions. So there I was, essentially an anthropologist on Mars, unable to convey to one of the best and brightest in the traditional ways of thinking that the whole idea of laboratory experiments was to evaluate mechanisms in an environment where *the Pareto optimal outcome was known by the experimental designer but not by the agents* so that performance comparisons could be made; that in the field such knowledge was never possible, and we had no criteria, other than internal theoretical properties such as incentive compatibility to judge the efficacy of the mechanism. He didn't get it; psychologically this testing procedure is not comprehensible if somehow your thinking has accustomed you to believe that allocation mechanisms require agents to have complete information, but not mechanism designers who presumably slipped through by assuming their agents were fully informed. In fact, with that worldview what is there to test in mechanism theory?

The issue of whether economics is an experimental science is moot among experimental economists who are, and should be, too busy having fun doing their work to reflect on the methodological implications of what they do. But when we do speak of methodology, as in comprehensive introductions to the field, what do we say? Quotations from impeccable sources will serve to introduce the concepts to be developed next. The first emphasizes that an important category of experimental work '… includes experiments designed to test the predictions of well articulated formal theories and to observe unpredicted regularities, in a controlled environment that allows these observations to be unambiguously interpreted in relation to the theory' (Kagel and Roth 1995, p. 22). Experimental economists strongly believe, I think, that this is our most powerful scientific defence of experimental methods: we ground our experimental inquiry in the firm bedrock of

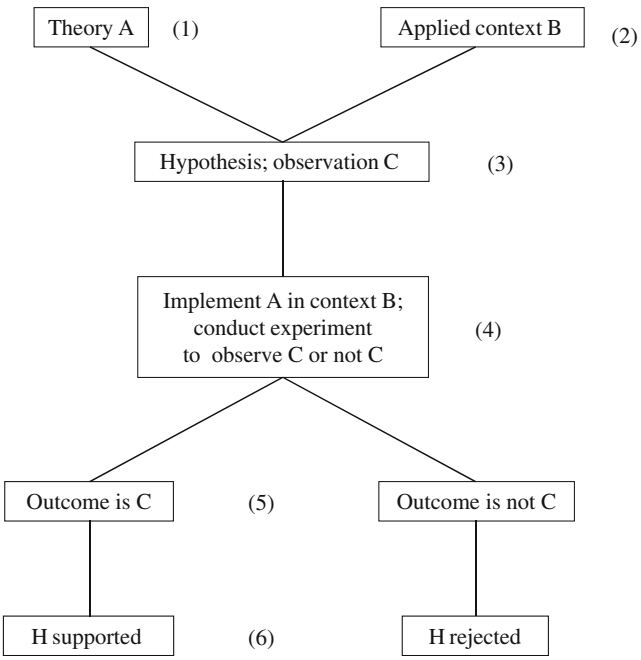
economic or game theory. A second crucial advantage, recognizing that field tests involve hazardous joint tests of multiple hypotheses, is the sentiment that 'Laboratory methods allow a dramatic reduction in the number of auxiliary hypotheses involved in examining a primary hypothesis' (Davis and Holt 1993, p. 16).

Hence the strongly held belief that, in the laboratory, we can test well- articulated theories, interpret the results unambiguously in terms of the theory, and do so with minimal, trivial or at least greatly reduced dependence on auxiliary hypothesis. This view and the idea that theories can be derived directly from observations are not unique to any science, but they are *illusions*. Fortunately, such illusions do not constitute a barrier to great scientific achievement because they appear to affect the rhetoric of science far more than its substance. Perhaps this is because the beliefs of scientists are important in reinforcing their commitment to discovery whether or not they are defensible. I cannot imagine that Newton would have been more accomplished if he had been methodologically more sophisticated.

What Is the Scientist's (qua Experimentalist) Image of What He Does?

The standard experimental paper within and without economics uses the following format in outline form: (1) state the theory; (2) implement it in a particular context (with 'suitable' motivation in economics); (3) summarize the implications in one or more testable hypotheses; (4) describe the experimental design; (5) present the data and results of the hypothesis tests; (6) conclude that the experiments either reject or fail to reject the theoretical hypotheses. This format is shown in Fig. 1. In the case in which we have two or more competing theories and corresponding hypotheses, the researcher offers a conclusion as to which one is supported by the data using some measure of statistical distance between the data and each of the predictive hypotheses, reporting which distance is statistically the shortest.

Suppes (1969; also see Mayo 1996, ch. 5) has observed that there exists a hierarchy of models behind the process in Fig. 1. The primary model or theory is contained in steps (1) and (2), which



Experimental Economics, Fig. 1 The scientist's image of scientific procedure

generate particular topical hypotheses that address primary qsts. Experimental models are contained in (3) and (4). These serve to link the primary theory with data. Finally, we have data models, steps (5) and (6), that link operations on raw data (not the raw data itself) to the testing of experimental hypotheses.

This process describes much of the rhetoric of science, and reflects the self-image of scientists, but it does not adequately articulate what scientists actually do. Furthermore, the rhetoric does not constitute a viable, defensible and coherent methodology. But what we actually do, I believe, is highly defensible and on the whole positively affects what we think we know from experiment. Implicitly, as experimentalists, we understand that every step, (1)–(6), in the above process is subject to judgments, learning from past experiments, our knowledge of protocols and technique, and to error. This is reflected in what we do as a professional community, if not in what we say about what we do in the standard scientific paper, or when we try to describe the science that we do.

As I have noted, the problem with the above image is known as the 'D–Q problem': experimental results always present a joint test of the theory (however well articulated, formally) that motivated the test, *and all the things you had to do to implement the test.* (For good discussions of the D–Q thesis and its relevance for experimental economics, see Soberg 2005, and Guala 2005, pp. 54–61 and *passim.* Soberg provides interesting theoretical results showing how the process of replication can be used, in the limit, to inductively eliminate clusters of alternative hypotheses and lend increasing weight to the conclusion that the theory itself is in doubt.) Thus, if theoretical hypothesis H is implemented with context specific auxiliary hypotheses required to make the test operational, A_1, A_2, . . ., A_n; then it is (H|A_1, A_2, . . ., A_n) that implies observation C. If you observe not-C, this can be because any of the antecedents (H; A_1, . . ., A_n) can represent what is falsified. Thus, the interpretation of observations in relation to a theoretical hypothesis is *inherently and inescapably ambiguous,* contrary to our accustomed rhetoric.

The reality of what we do, and indeed must do, is implied by the truth that 'No theory is or can be killed by an observation. Theories can always be rescued by auxiliary hypotheses' (Lakatos 1978, vol. 1, p. 34).

A D–Q Example from Physics

Here is a historical example from physics: in 1905 Kaufmann (cited in Fölsing 1997, p. 205), a very accomplished experimentalist (in 1902 he showed that the mass of an electron is increased by its velocity!), published a paper 'falsifying' Einstein's special theory of relativity the same year in which the latter was published (Einstein 1905). Subsequently, Einstein (1907) in a review paper reproduced Kaufmann's Figure 2, commenting that

> The little crosses above the [Kaufmann's] curve indicate the curve calculated according to the theory of relativity. In view of the difficulties involved in the experiment one would be inclined to consider the agreement as satisfactory. However, the deviations are systematic and considerably beyond the limits of error of Kaufmann's experiment. That the calculations of Mr. Kaufmann are error-free is shown by the fact that, using another method of calculation, Mr. Planck arrived at results that are in full agreement with those of Mr. Kaufmann. Only after a more diverse body of observations becomes available will it be possible to decide with confidence whether the systematic deviations are due to a not yet recognized source of errors or to the circumstance that the foundations of the theory of relativity do not correspond to the facts. (Einstein 1907, p. 283.

Kaufmann was testing the hypothesis that (H|A) implies C, where H was an implication of special relativity theory, A was the auxiliary hypothesis that his context-specific test and enabling experimental apparatus was without 'a not yet recognized source of errors', C was the curve indicated by the 'little crosses', and not-C was Kaufmann's empirical curve (Einstein 1907, Figure 2, p. 283).

Einstein's comment in effect says that either of the antecedents (H|A) represents what is falsified by Kaufmann's results. Others, such as Planck and Kaufmann himself, however, acknowledged that the observation might conceivably be in error

(Fölsing 1997, p. 205). Such acknowledgements are not unusual in the scientific community, which means that scientists informally recognize the D–Q problem as it arises in particular contexts, that it is part of the scientific conversation and that they seek solutions even if this modus operandi is not part of their rhetoric. Thus, in less than a year, Bucherer (see Fölsing 1997, p. 207) showed that there indeed had been a 'problem' with Kaufmann's experiments and proceeded to obtain new results supporting Einstein's theory.

There is an important lesson in this example if we develop it a little more fully. Suppose Bucherer's experiments *had not changed Kaufmann's results enough to change the conclusion* (There is never a shortage of claims that a given experimental result may be in doubt: see Mayo 1996, for the imaginative arguments proffered by the Newtonians in response to Eddington's eclipse observations). Then Einstein could still have argued that there may be 'a not yet recognized source of errors'. If so, the implication is that H is not falsifiable, for the same argument can be made after each new test in which the new results are outside the range of error for the apparatus! Recall that the deviations were alleged to be 'considerably beyond the limits of error of Kaufmann's experiment'. But here 'error' is used in the sense of internal variations arising from the apparatus and procedure, not in the sense that there is a problem with the apparatus or procedure itself. We can go still further in explicating the problem of testing H conditional on any A. The key is to note in this example the strong dependence of any test outcome on the *state of experimental knowledge:* Bucherer found a way to 'improve' Kaufmann's experimental technique so as to rescue Einstein's 'prediction'. But the predictive content of H (and therefore of the special theory) was inextricably bound up with A. Einstein's theory did not embrace any of the elements of Kaufmann's (or Bucherer's) *apparatus:* A is based on experimental knowledge of testing procedures and operations in the physics laboratory, and has nothing to do with the theory of relativity, a separate and distinct body of theoretically coherent knowledge.

A Proposition and An Economics Example

Here is the most common casual empiricist objection to economics experiments: the payoffs are too small. This objection is one of several principal issues in a target article by Hertwig and Ortmann (2001), with comments by 34 experimental psychologists and economists. This objection sometimes is packaged with an elaboration to the effect that economic or game theory is about situations that involve large stakes, and you 'can't' study these in the laboratory (Actually, of course, you can, but funding is not readily available).

Suppose, therefore that we have the following:

H (from theory): subjects will choose the equilibrium outcome (for example, Nash or subgame perfect).

A (auxiliary hypothesis): payoffs are adequate to motivate subjects.

Proposition 2 Suppose a specific rigorous test rejects (H|A), and someone (say, T), protests that what must be rejected is A not H. Let E replicate the original experiment with an n-fold increase in payoffs. There are only two outcomes and corresponding inpts, neither of which is comforting to the rhetorical image of science as conducting falsification tests of predictive hypotheses:

1. The test outcome is negative. Then T can imagine a still larger payoff multiple $N > n$, and still argue for rejecting A not H. But this implies that H *cannot be falsified*.
2. After repeated increases in payoffs, the test outcome is positive for some N ^ n*. Then H has *no predictive content.* E, with no guidelines from the theory, has searched for and discovered an empirical payoff multiple, n^*, that 'confirms' the theory, but n^* is an *extra theoretical property of considerations outside H and the theory* that was being tested. Finding this multiple is not something for T or E to crow about, but rather an event that should send T or E back to his desk. The theory is inadequately specified to embrace the observations from all the experiments.

Proposition 2 holds independent of any of the following considerations:

- how well articulated, rigorous or formal the theory is; game theory in no sense constitutes an exception;
- how effective the experimental design is in reducing the number of auxiliary hypotheses – it only takes one to create the problem; and
- the character or nature of the auxiliary hypothesis – *A* can be anything not contained in the theory.

In experimental economics, reward adequacy is just one of a standard litany of objections to experiments in general and to many experiments in particular. Here are three additional categories in this litany:

1. *Subject sophistication.* The standard claim is that undergraduates are not sophisticated enough. They are not out there participating in the 'real world'. In the 'real world' where the stakes are large, such as in the FCC spectrum rights auctions, bidders use game theorists as consultants (Banks et al. 2003) (For an investigation of the hypothesis that undergraduates are insufficiently sophisticated, see McCabe and Smith 2000, who report a comparison of undergraduate and graduate students, and these with economics faculty, in a two-person trust game. The first two groups were indistinguishable, and both earned more than the faculty because with greater frequency they risked defection in offering to cooperate, as against opting for the subgame perfect outcome).
2. *Subjects need an opportunity to learn*. This is a common response from both experimentalists and theorists when you report the results of single play games in which 'too many' people cooperate. The usual proposed 'fix' is to do repeat single protocols in which distinct pairs play on each trial, and apply a model of learning to play non-cooperatively (See McCabe et al. 1998, for a trust game with the option of punishing defection in which support for the cooperative outcome does not decrease in repeat single relative to single play across trials, and therefore subjects *do not* 'learn' to play non-cooperatively). But there are many unanswered qsts implicit in this auxiliary hypothesis: since repeat single protocols require large groups of subjects (20 subjects to produce a 10-trial sequence), have any of these games been run long enough to measure adequately the extent of learning? In single play two-person anonymous trust games data have been reported showing that group size matters; that is, it makes a difference whether you are running 12 subjects (6 pairs) or 24 subjects (12 pairs) simultaneously in a session (Burnham, McCabe and Smith 2000). Also, in the larger groups pairs were found to be less trusting than in the small groups – perhaps not too surprising. But in repeat single games, in which a game is repeated with distinct pairs of subjects on each repetition, larger groups are needed for longer trial sequences. Hence, learning and group size as auxiliary hypotheses loses independence, and we have knotty new problems of complex joint hypothesis testing. The techniques, procedures and protocol tests we fashion for solving such problems are the sources of our experimental knowledge. All testing depends on, and is limited by, the state of that experimental knowledge at any given time. Over time it expands incrementally in the design problem-solving context of particular new testing challenges. This is a community development enterprise that is largely outside individual conscious awareness, but an integral part of the sociality of scientific change.
3. *Instructions are adequate* (or decisions are robust with respect to instructions, and so on). What does it mean for instruction to be adequate? Clear? If so, clear about what? What does it mean to say that subjects understand the instructions? Usually this is interpreted to mean that they can perform the task, which is judged by how they answer qsts about what they are to do. In two-person interactions, instructions often matter so much that they must be considered a (powerful) treatment (Thus, Hertwig and Ortmann 2001, section 2, argue that scripts – instructions – are important for replication, and that 'ad-libbing'

should be avoided). Instructions can be important because they define context, and context matters. Ultimatum and dictator game experiments yield statistically and economically significant differences in results due to differing instructions and protocols (Hoffman et al. 1994; Hoffman et al. 1996).

Positive Economics: Judge Theories by Their Predictions Not Their Assumptions

There is a methodological perspective associated with Milton Friedman (1953), which fails to provide an adequate foundation for experimental (field or laboratory) science, but which influenced economists for decades and still has some currency. Friedman's proposition is that the truth value of a theory should be judged by its testable and tested predictions not by its assumptions. This proposition is deficient for at least three reasons:

1. If a theory fails a test, we should ask why, not always just move on to seek funding for a different new project; obviously, one or more of its assumptions may be wrong, and it behoves us to design experiments that will probe conjectures about which assumptions failed. Thus, if first price auction theory fails a test is it a consequence of the failure of one of the axioms of expected utility theory, for example, the compound lottery axiom? If a subgame perfect equilibrium prediction fails, does the theory fail because the subjects do not satisfy the assumption that the agents choose dominant strategies? Or did the subjects fail to use backward induction? Or was it none of the above because the theory was irrelevant to how some motivated agents solve such problems in a world of bilateral (or multilateral) reciprocity in social exchange? When a theory fails there is no more important question to ask than what it is about the theory that has failed.
2. Theories may have the if-and-only if property that one (or more) of their 'assumptions' can be derived as implication(s) from one (or more) of their 'results'. These cases if trivial lead to the reversible property of testing illustrated above for risk-neutral agents bidding in first price auctions with linear density functions on value.
3. If a theory passes one or more tests, this need not be because its assumptions are correct. A subject may choose the subgame perfect equilibrium because she believes she is paired with a person that is not trustworthy, and not because she always chooses dominant strategies, or assumes that others always so choose or that this is common knowledge. This is why you are not done when a theory is corroborated by a test. You have examined only one point in the parameter space of payoffs, subjects, tree structure, and so on. Your results may be a freak accident of nature due to a complex of suitabilities or in any case may have other explanations.

In View of Proposition 2, What Are Experimentalists and Theorists to Do?

Consider first the example in which we have a possible failure of A: rewards are adequate to motivate subjects. Experimentalists should do what comes naturally, namely, do new experiments that increase rewards and/or lower decision costs by simplifying experiment procedures. The literature is full of examples (for surveys and inpts see Smith and Walker 1993; Camerer and Hogarth 1999; Holt and Laury 2001; Harrison et al. 2005).

Theorists should ask whether the theory is extendable to include *A*, so that the effect of payoffs is explicitly modelled. It is something of a minor scandal when economists – whose models predict the optimal outcome independently of payoff levels, and however gently rounded the payoff in the neighbourhood of the optimum is – object to an experiment because the payoffs are inadequate. What is adequate? Why are payoff inadequacies a complaint rather than a spur to better and more relevant modelling of payoffs? A step toward modelling both H and A (as payoffs) is provided in Smith and Szidarovszky (2004). Economic intuition tells us that payoffs should matter. But if they do, it must mean that some cost, which is impeding equilibrium, is being overcome at the margin by higher payoffs.

The natural psychological assumption is that there are cognitive costs of getting the best outcome, and more reward enables higher cognitive cost to be incurred to increase net return.

Generally, both groups must be aware that for any H, any A and any experiment, one can say that if the outcome of the experiment rejects (H|A), then both should assume that either H or A may be false, which is an obvious corollary to Proposition 1. This was Einstein's position concerning the Kaufmann results, and was the correct position, whatever later tests might show. After every test, either the theory (leading to H) is in error or the circumstances of the test, A, is in error.

The experimentalist has much to do, but primarily more experiments, which is precisely what experimentalists do in response to the many conjectures about what is wrong with the experiment – re-examine the instructions, payoffs, subjects, anything and everything the experimentalist did to formulate the test.

The theorist should also ask, especially if further tests continue to reject (H|A), whether the auxiliary hypothesis can be incorporated into the theory.

If the outcome fails to reject (H|A), the experimentalist should escalate the severity of the test. At some point does H fail? This identifies the limits of the theory's validity, and gives the theorist clues for modifying the theory.

Experimental Knowledge Drives Our Methods

Philosophers have written eloquently and argued intently over the implications of D–Q and related issues for the interpretation of science. Popper tried to demarcate science from pseudoscience with the basic rule requiring the scientist to specify in advance the conditions under which he will give up his theory. This is a variation on the idea of a so-called 'crucial experiment', a concept which cannot withstand examination (Lakatos 1978, vol. 1, pp. 3–4, 146–8), as is evident from our Proposition 2.

The failure of all programmes attempting to articulate a defensible rational science of scientific method has bred postmodern negative reactions to science and scientists. These exercises and controversies make fascinating reading, and provide a window on the social context of science, but I believe they miss the essence of what is most important in the working lives of all practitioners. Popper was wrong in thinking he could demarcate science from pseudoscience by an exercise in logic, but that does not imply that the Popperian falsification rule failed as a milestone contribution to the scientific conversation; nor does it mean that 'anything goes' (Feyeraband 1975). Rather, what one can say is much less open ended: anything goes only in so far as what can be concluded about *constructive rationality in science.* But the scientific enterprise is also about *ecological rationality in science,* which is about discovery, about probes into Max Born's 'jungle', about thinking outside the box and, as I shall argue below, about the technology of observation in science that renders obsolete long-standing D–Q problems while introducing new ones for a time.

You do not have to know anything about D–Q and statements like Proposition 2 to appreciate that the results of an experiment nearly always suggest new qsts precisely because the interpretation of results in terms of the theory are commonly *ambiguous.* This ambiguity is reflected in the discussion whenever new results are presented at seminars and conferences. Without ambiguity there would be nothing to discuss. What is the response to this ambiguity? Invariably, if it is a matter of consequence, experimentalists design new experiments with the intention of confronting the issues in the controversy, and in the conflicting views that have arisen in interpreting the previous results. This leads to new experimental knowledge of how results are influenced, or not, by changes in procedures, context, instructions and control protocols. The new knowledge may include new techniques that have application to areas other than the initiating circumstance. This ecological process is driven by the D–Q problem, but practitioners need have no knowledge of the philosophy of science literature to take the right next steps, subject to error, in the laboratory.

This is because the theory or primary model that motivates the qsts tells you nothing definitive

or even very useful about precisely how to construct tests. Tests are based on extra theoretical intuition, conjectures, and experiential knowledge of procedures. The context, subjects, instructions, parameterization, and so on are determined outside the theory, and their evolution constitutes the experimental knowledge that defines our methodology. The forms taken by the totality of individual research testing programmes cannot be accurately described in terms of the rhetoric of falsification, no matter how much we speak of the need for theories to be falsifiable, stating in advance the conditions under which the theory will be rejected, the tests discriminating or 'crucial' and the results robust.

Whenever negative experimental results threaten perceived important new theoretical tests, the resulting controversies galvanize experimentalists into a search for different or better tests – tests that examine the robustness of the original results. Hence, Kaufmann's experimental apparatus was greatly improved by Bucherer, although there was no question about Kaufmann's skill and competence in laboratory technique. The point is that with escalated personal incentives, and a fresh perspective, scientists found improved techniques. This scenario is common as new challenges bring forth renewed effort. This process generates the constantly changing body of experimental and observational knowledge whose insights in solving one problem often carry over to applications in many others.

Just as often experimental knowledge is generated from curiosity about the properties of phenomena that we observe long before a body of theory exists that deals specifically with the phenomenon at issue. An example in experimental economics is the continuous double auction trading mechanism (Smith 1962, 2003).

An example from physics is Brownian motion, discovered by the botanist Robert Brown in 1827, who first observed the unpredictable motion of small particles suspended in a fluid. This motion is what keeps them from sinking under gravity. This was 78 years before Einstein's famous paper (one of three) in 1905 developed the molecular kinetic theory that was able to account for it, although he did not know that the applicable observations of Brownian motion were already long familiar (see Mayo 1996, ch. 7, for references and details). The long 'inquiry into the cause of Brownian motion has been a story of hundreds of experiments . . . [testing hypotheses attributing the motion]. . .either to the nature of the particle studied or to the various factors external to the liquid medium . . . '(Mayo 1996, pp. 217–18). The essential point is 'that these early experiments on the possible cause of Brownian motion were not testing any full-fledged theories. Indeed it was not yet known whether Brownian motion would turn out to be a problem in chemistry, biology, physics, or something else. Nevertheless, a lot of information was turned up and put to good use by those later researchers who studied their Brownian motion experimental kits' (Mayo 1996, p. 240). The problem was finally solved by drawing on the extensive bag of experimental tricks, tools and past mistakes that constitute 'a log of the extant experimental knowledge of the phenomena in qst' (1996, p. 240).

Again, 'the infinitely many alternatives really fall into a few categories. Experimental methods (for answering new qsts) coupled with experimental knowledge (for using techniques and information already learned) enable local qsts to be split off and answered' (Mayo 1996, p. 242).

The bottom line is that good-enough solutions emerge to the baffling infinity of possibilities, as new measuring systems emerge, experimental toolkits are updated, and understanding is sharpened. This bottom line also goes far towards writing the history of experimental economics and its many detailed encounters with data, and the inevitable ambiguity of subsequent inpt. And in most cases the jury remains in session on whether we are dealing with a problem in psychology (perception), economics (opportunity cost and strategy), social psychology (equality, equity or reciprocity), neuroscience (functional imaging and brain modeling) or all of the above. So be it.

The Machine Builders

Mayo's (1996) discussion and examples of experimental knowledge leave unexamined the

question of how technology affects the experimentalist's toolkit. The heroes of science are neither the theorists nor the experimentalists but the unsung tinkers, mechanics, inventors and engineers who create the new generation of machines that make obsolete yesterday's observations and heated arguments over whether it is T or A that has been falsified. Scientists, of course, are sometimes a part of this creative destruction, but what is remembered in academic recognition is the new scientific knowledge they created, not the instruments they invented that made possible the new knowledge. Michael Faraday, 'one of the greatest physicists of all time' (Segrè 1984, p. 134), had no formal scientific education. He was a bookbinder, who had the habit of reading the books that he bound. He was preeminently a tinker for whom 'some pieces of wood, some wire and some pieces of iron seemed to suffice him for making the greatest discoveries' (quoted from a letter by Helmholz in Segrè 1984, p. 140). Yet he revolutionized how physicists thought about electromagnetic phenomena, invented the concept of lines of force (fields), and inspired Maxwell's theoretical contributions. 'He writes many times that he must experience new phenomena by repeating the experiments, and that reading is totally insufficient for him' (Segrè 1984, p. 141). This is what I mean, herein, when I use the term 'experimental knowledge'. It is 'can do' knowledge acquired by trial, error and discovery. And it is what Mayo (1996) is talking about. It is also why doing experiments changed the way I thought about economics.

Technology and Science

With the first moon landing, theories of the origin and composition of our lunar satellite, contingent on the state of existing indirect evidence, were upstaged by direct observation; the first Saturn probe sent theorists back to their desks and computers to re-evaluate her mysterious rings, whose layered richness had not been anticipated. Similar experiences have followed the results of ice core sampling in Greenland, and instrumentation for mapping the genome of any species. Galileo's primitive telescope opened a startling window on the solar system, as do Roger Angel's multiple mirror light-gathering machines (created under the Arizona football stadium) that open the mind to a view of the structure of the universe. (For a brief summary of the impact of past, current, and likely future effects of rapid change in optical and infrared (terrestrial and space) telescopes on astronomy see Angel 2001.) The technique of tree ring dating, invented by an Arizona astronomer, has revolutionized the interpretation of archeological data from the last 5,000 years.

Yesterday's reductionisms, shunned by mainstream 'practical' scientists, create the demand for new deeper observations, and hence for the machines that can deliver answers to entirely new qsts. Each new machine – microscope, telescope, Teletype, computer, the Internet, fMRI imaging – changes the way teams of users think about their science. The host of auxiliary hypotheses needed to give meaning to theory in the context of remote and indirect observations (inferring the structure of Saturn's ring from earth-based telescopes) are suddenly made irrelevant by deep and more direct observations of the underlying phenomena (fly by computer-enhanced photos). It's the machines that drive the new theory, hypotheses, and testing programmes that take you from atoms, to protons, to quarks. Yet with each new advance comes a blizzard of auxiliary hypotheses, all handcuffed to new theory, giving expression to new controversies seeking to rescue T and reject A, or to accept A and reject T.

Experimental Economics and Computer/Communication Technology

In 1976 when Arlington Williams (1980) created the first electronic double-auction software program for market experiments, all of us thought we were simply making it easier to run experiments, collect more accurate data, observe longer time series, facilitate data analysis, and so on. What were computerized were the procedures, recording and accounting that heretofore had been done manually. No one was anticipating how this tool might impact on and change the way we thought

about the nature of doing experiments. But with each new electronic experiment we were 'learning' (affected by, but without conscious awareness of) the fact that traders could be matched at essentially zero cost, that the set of feasible rules that could be considered was no longer restricted by costly forms of implementation and monitoring, that vastly larger message spaces could be accommodated, and that optimization algorithms could now be applied to the messages to define new electronic market forms for trading energy, air emission permits, water and other network industries. In short, the transaction cost of running experimental markets became minuscule in comparison with the pre-electronic days, and this opened up new directions that previously had been unthinkable.

This quickly led to the concept of smart computer-assisted markets, which appeared in the early 1980s (Rassenti 1981; Rassenti et al. 1982), extended conceptually to electric power and gas pipelines in the late 1980s (Rassenti and Smith 1986; McCabe et al. 1989), with practical applications to electric power networks and the trading of emission permits across time and regions in the 1990s (Rassenti et al. 2002). These developments continue as major new efforts in which the laboratory is used as a test bed for measuring, modifying, and further testing the performance characteristics of new institutional forms.

What is called e-commerce has spawned a rush to reproduce on the Internet the auction, retailing and other trading systems people know about from past experience. But the new experience of being able to match traders at practically zero cost is sure to change how people think about trade and commerce, and ultimately this will change the very nature of trading institutions. In the short run, of course, efforts to protect existing institutions will spawn efforts to shield them from entry by deliberately introducing cost barriers, but in the long run these efforts will be increasingly uneconomical.

Neuroscience carries the vision of changing the experimental study of individual, two-person interactive and market decision making. The neural correlates of decision making, how it is affected by rewards, cognitive constraints, working memory, institutions, repeat experience and a host of factors that in the past we could neither control or observe can in the future be expected to become an integral part of the way we think about and model decision making. Models of decision, now driven by game and utility theory, and based on trivial, patently false, models of mind, must take account of new models of cognitive, calculation and memory properties of mental function that are accessible to more direct observational inpt. Game-theoretic models assume consciously calculating, rational mental processes, but models of mind include non-self-aware processes just as accessible to neural brain imaging as the conscious. For the first time we may be able to give some observational content to the vague and slippery idea of 'bounded rationality' (see Camerer et al. 2005).

Conclusion

In principle the D–Q problem is a barrier to any defensible notion of a rational science that selects theories by a logical process of confrontation with scientific evidence. This is cause for joy not despair. Think how dull a life of science would be if, once we were trained, all we had to do was to turn on the threshing machine of science, feed it the facts, send its output to the printer, and run it through the formulas for writing a scientific paper.

As I see it, there is no rationally constructed science of scientific method. The attempt to do it has led to important insights and understanding, and has been a valuable exercise. But all construction must ultimately pass ecological or 'fitness' tests based on the totality of our experience. Control is of course important; it is why we do laboratory and field experiments. But control is always limited in scope, and above all the rhetoric of control should not restrict the examination and reexamination of our own assumptions, both in the theory and in its testing, or limit our capacity to think outside the professional box. We do this in the reality underneath our rhetoric because we cannot help it, so much is it part of our deep human sociality and the workings of our social brains.

See Also

► Experimental Economics, History Of

Bibliography

Angel, R. 2001. Future optical and infrared telescopes. *Nature Insights* 409: 427–430.

Banks, J., M. Olson, D. Porter, S. Rassenti, and V. Smith. 2003. Theory, experiment and the federal communications commission spectrum auctions. *Journal of Economic Behavior & Organization* 51: 303–350.

Born, M. 1943. *Experiment and theory in physics*. Cambridge: Cambridge University Press.

Burnham, T., K. McCabe, and V. Smith. 2000. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization* 43: 57–73.

Camerer, C.F., and R.M. Hogarth. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19: 7–42.

Camerer, C., G. Loewenstein, and D. Prelec. 2005. Neuroeconomics: How neuroeconomics can inform economics. *Journal of Economic Literature* 43: 9–64.

Cox, J., V. Smith, and J. Walker. 1988. Theory and individual behavior of first price auctions. *Journal of Risk and Uncertainty* 1: 61–99. Reprinted in Smith (1991).

Davis, D., and C. Holt. 1993. *Experimental economics*. Princeton: Princeton University Press.

Davis, D., and A. Williams. 1991. The Hayek hypothesis in experimental auctions: Institutional effects and market power. *Economic Inquiry* 29: 261–274.

Einstein, A. 1905. On the electrodynamics of moving bodies. In *The collected papers of Albert Einstein,* vol. 2, trans. A. Beck, Princeton: Princeton University Press, 1989.

Einstein, A. 1907. On the relativity principle and the conclusions drawn from it. In *The collected papers of Albert Einstein,* vol. 2, trans. A. Beck. Princeton: Princeton University Press, 1989.

Einstein, A. 1934. *The world as I see it*. New York: Covici Friede Publishers.

Feyeraband, P. 1975. *Against method*. London: Versa.

Fölsing, A. 1997. *Albert Einstein*. New York: Viking.

Friedman, M. 1953. *Essays in positive economics*. Chicago: University of Chicago Press.

Grandin, T., and C. Johnson. 2005. *Animals in translation*. New York: Scriber.

Guala, F. 2005. *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Harrison, G., J. McInnes, and L. Rutstroem. 2005. Risk aversion and incentive effects: Comment. *American Economic Review* 95: 897–901.

Hayek, F.A. 1942. Scientism and the study of society. *Economica;* reprinted in *The Counter-Revolution in Science.* Indianapolis: Liberty Press, 1979.

Hayek, F.A. 1945. The use of knowledge in society. *American Economic Review* 35: 519–530.

Hayek, F.A. 1967. *Studies in philosophy, politics and economics*. London: Routledge & Kegan Paul.

Hayek, F.A. 1978. Competition as a discovery procedure. In *New studies in philosophy, politics, economics, and the history.* Chicago: Chicago University Press. Reprinted in *The Essence of Hayek.* Stanford: Hoover Institution Press, 1984.

Hertwig, R., and A. Ortmann. 2001. Experimental practices in economics. *Behavioral and Brain Sciences* 24: 383–451.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith. 1994. Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior* 7: 346–380.

Hoffman, E., K. McCabe, and V. Smith. 1996. On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory* 25: 289–301.

Holt, C. 1989. The exercise of market power in experiments. *Journal of Law and Economics* 32S: 107–130.

Holt, C., and S. Laury. 2001. Varying the scale of financial incentives under real and hypothetical conditions. *Behavioral and Brain Sciences* 24: 417–418.

Holt, C., L. Langan, and A. Villamil. 1986. Market power in oral double auction experiments. *Economic Inquiry* 24: 107–123.

Kagel, J., and A. Roth. 1995. *Handbook of experimental economics*. Princeton: Princeton University Press.

Kahneman, D., J. Knetsch, and R. Thaler. 1986. Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review* 76: 728–741.

Lakatos, I. 1978. *The methodology of scientific research programmers*. Vol. 2. Cambridge: Cambridge University Press.

Leamer, E. 1978. *Specification searches*. New York: Wiley.

Mayo, D. 1996. *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

McCabe, K., and V. Smith. 2000. A comparison of naïve and sophisticated subject behavior with game theoretic prediction. *Proceedings National Academy of Sciences* 97: 3777–3781.

McCabe, K., S. Rassenti, and V.L. Smith. 1989. Designing smart computer assisted markets for gas networks. *European Journal of Political Economy* 5: 259–283.

McCabe, K., Rassenti, S. and Smith, V.L. 1998. Reciprocity, trust and payoff privacy in extensive form bargaining. *Games and Economic Behavior* 24: 10–24. Reprinted in Smith (2000).

Northrup, F.S.C. 1969. Einstein's conception of science. In *Albert Einstein: Philosopher–scientist*, ed. P.A. Schilpp. LaSalle: Open Court.

Pirsig, R.M. 1981. *Zen and the art of motorcycle maintenance*. New York: Banton Books.

Rassenti, S. 1981. O-1 decision problems with multiple resource constrains: Algorithms and applications. Ph.D. thesis, University of Arizona.

Rassenti, S., and V. Smith. 1986. Electric utility deregulation. In *Pricing electric gas and telecommunication*

services. Washington, DC: Institute for the Study of Regulation.
Rassenti, S., V. Smith, and R. Bulfin. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13: 402–417.
Rassenti, S., V. Smith, and B. Wilson. 2002. Using experiments to inform the privatization/deregulation movement in electricity. *Cato Journal* 21: 515–544.
Robinson, J. 1979. What are the qsts? *Journal of Economic Literature* 15: 1318–1339.
Samuelson, P. 1966. Intertemporal price equilibrium: A prologue to the theory of speculation. In *The collected papers of Paul A. Samuelson*, ed. J. Stiglitz, vol. 2. Cambridge, MA: MIT Press.
Samuelson, P., and W. Nordhaus. 1985. *Economics*. New York: McGraw-Hill.
Segrè, E. 1984. *From falling bodies to radio waves*. New York: Freeman.
Shubik, M. 1959. *Strategy and market structure*. New York: Wiley.
Smith, V.L. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70: 111–137.
Smith, V.L. 1965. Experimental auction markets and the Walrasian hypothesis. *Journal of Political Economy* 73: 387–393.
Smith, V.L. 1976. Experimental economics: Induced value theory. *American Economic Review Proceedings* 66: 274–279 .Reprinted in Smith (1980; 1991)
Smith, V.L. 1980. *Evaluation of econometric models*. New York: Academic Press.
Smith, V.L. 1991. *Papers in experimental economics*. New York: Cambridge University Press.
Smith, V.L. 2000. *Bargaining and market behavior. Essays in experimental economics*. New York: Cambridge University Press.
Smith, V. 2003. Constructivist and ecological rationality in economics. *American Economic Review* 93: 465–508.
Smith, V.L., and F. Szidarovszky. 2004. Monetary rewards and decision cost in strategic interactions. In *Models of a man: Essays in memory of Herbert A. Simon*, ed. M. Augier and J. March. Cambridge, MA: MIT Press.
Smith, V., and J. Walker. 1993. Monetary rewards and decision cost in experimental economics. *Economic Inquiry* 31: 245–261 .Best Article Award, Western Economic Association, 1993
Smith, V., K. McCabe, and S. Rassenti. 1991. Lakatos and experimental economics. In *Appraising economic theories*, ed. N. De Marchi and M. Blaug. London: Edward Elgar.
Soberg, M. 2005. The Duhem–Quine thesis and experimental economics: A reinpt. *Journal of Economic Methodology* 12: 581–597.
Stigler, G. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65: 1–17.
Suppes, P. 1969. Models of data. In *Studies in the methodology and foundations of science*. Dordrecht: Reidel.
Williams, A. 1980. Computerized double-auction markets: Some initial experimental results. *Journal of Business* 53: 235–257.

Experimental Economics, History of

Francesco Guala

Abstract

Contemporary experimental economics was born in the 1950s from the combination of the experimental method used in psychology and new developments in economic theory. Early experimental studies of bargaining behaviour, social dilemmas, individual decision making and market institutions were followed by a long period of underground growth, until the booming of the field in the 1980s and 1990s.

Keywords

Allais paradox; Allais, M; Auctions (experiments); Bargaining; Behavioural economics; Cardinal utility; Convergence; Decision theory; Expected utility; Experimental economics; Experimental economics, history of; Friedman, M; Game theory; Gaming; Kahneman, D; Learning; Mathematics and economics; Mechanism design; Mill, J. S; Models; Other-regarding preferences; Plott, C; Positive economics; Preference reversals; Prisoner's dilemma; Public goods; Ramsey, F; Savage, L; Selten, R; Siegel, S; Simon, H; Smith, V; Stanford Value Project; Statistics and economics; Subgame perfection; Subjective probability; Tversky, A; Ultimatum game; Utility measurement; Von Neumann and Morgenstern

JEL Classifications

B4

Experimental economics has experienced one of the most stunning methodological revolutions in the history of science. In just a few decades, economics has been transformed from a discipline where the experimental method was considered impractical, ineffective and largely irrelevant to

one where some of the most exciting advancements are driven by laboratory data.

Like many other new developments in the social sciences during the second half of the 20th century, experimental economics is largely a by-product of the combination of massive investments in science, a fertile intellectual culture and socio-political conditions in the 1940s and 1950s in the United States. Although it is possible in principle to identify earlier experimental or proto-experimental work being done in economics and psychology (see Roth 1995), there is hardly any direct intellectual, personal or institutional continuity between these isolated episodes and today's fully institutionalized experimental programme.

A proper history of experimental economics is yet to be written, and one challenge faced by historians of the discipline is its strikingly interdisciplinary character. The rise of experimental economics takes the form of several, partly independent and partly intertwined threads that can be brought under a single coherent narrative only with difficulty. It is partly for this reason that most of the existing historical literature consists of personal recollections or reconstructions of individual trajectories rather than of a collective enterprise. It is possible, however, to identify some key moments and achievements that have helped to establish experimentation as a legitimate method of investigation in economics.

Historical Background and Early Years

The traditional view of economics as a primarily non-experimental science was outlined in the methodological writings of 19th-century economists. Mill (1836, p. 124), for example, identifies several practical obstacles to the use of the experimental method, in particular the impossibility of controlling key economic variables and of keeping background conditions fixed so that the effect of manipulating each cause in isolation can be checked. This was Mill's main justification for adopting the so-called 'a priori deductive' method, a mix of introspection and theoretical reasoning, to determine what an idealized *homo oeconomicus* would do in given circumstances. Despite various changes in economists' methodological rhetoric and practice, it took a century and a half for philosophical scepticism towards experimentation to fade away.

Like many methodological revolutions in science, the experimental turn in economics was primarily made possible not by a change in philosophical perspective but by a number of innovations at the level of scientific practice and theoretical commitment. At a very general level, in the middle of the 20th century economics was in the process of becoming a 'tool-based' science (Morgan 2003): from the old, discursive 'moral science' of political economy, it was changing into a discipline where models, statistics and mathematics played the role both of instruments and, crucially, of *objects* of investigation. During this conceptual revolution economists came to accept that the path towards the understanding of a real-world economy might have to go through the detailed analysis of several tools that had apparently only a vague resemblance to the final target of investigation. Theoretical models and computer simulations entered the economists' toolkit first, with laboratory experiments following shortly after.

The birth of experimental economics owes much to the publication of von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (1944) and to the subsequent developments of game and decision theory. Although game theory is often seen primarily as a contribution to the theoretical corpus of economics, this was not how it was perceived at the time. Von Neumann and Morgenstern's work initially found fertile ground in a community of scientists devoted to the simultaneous development of a great variety of approaches and research methods and interested in their application to solve scientific, policy, and management problems across the disciplinary boundaries – from conflict resolution in international relations to group psychology, cybernetics, and the organization of the firm, to name just a few.

'Gaming' – playing game-theoretic problems for real – was common practice in the mathematical community at Princeton in the 1940s and

1950s, and quickly spread elsewhere as game theory increased in popularity. This practice did not involve sophisticated experimental design, but was conceived mainly as a useful way of illustrating game theoretic puzzles, as well as a check on abstract speculation and a guide to the theoretician's intuitions. Traces of this attitude can be found in the writings of some pioneers in game theory in the 1950s, who explicitly advocated a combination of formal theorizing and empirical evidence of various kinds, and engaged in (mostly casual) forms of experimenting to back up their theoretical claims (see, for example, Schelling 1960; Shubik 1960).

The first event devoted specifically to 'The Design of Experiments in Decision Processes' was a 1952 two-month seminar sponsored by the Ford Foundation, organized in Santa Monica by a group of researchers at the University of Michigan. The seminar's location was intended to facilitate the participation of members of the RAND Corporation, a think tank sponsored by the US Air Force, where among others Merrill Flood was conducting game-theoretic experiments (including famously the first Prisoner's Dilemma experiments). It is difficult to assess at all precisely the role of the Santa Monica seminar in the birth of experimental economics because, apart from an important minority, most of the published papers (in Thrall et al. 1954) are theoretical rather than experimental in character. Several later protagonists, however, first became familiar with the idea of experiments in economics through the Santa Monica seminar, which therefore functioned as a catalyst in various indirect ways (see Smith 1992).

The most extensive experimental projects of the 1950s were pursued at Penn State, Michigan, and Stanford. In collaboration with Lawrence Fouraker, the psychologist Sidney Siegel conducted a systematic investigation of bargaining behaviour at Pennsylvania State University, trying to combine what he took to be the most advanced aspects of economics (the theory) and psychology (the experimental method). The project came to an abrupt end with Siegel's death in 1961, but the resulting book (Siegel and Fouraker 1960) won the American Academy of Arts and Sciences best monograph prize. Siegel and Fouraker's experiments focused on several aspects of bargaining behaviour, but are particularly significant for the systematic study of variations in the monetary payoffs and in the information made available to the subjects. Interestingly, this research project was rather disconnected from current developments in axiomatic bargaining theory, focusing instead on testing various hypotheses from the psychological literature. 'Level of aspiration theory' emerged eventually as the best predictor of bargaining behaviour.

From the point of view of experimental design, Siegel is often credited with being the first experimenter to highlight the importance of using real incentives to motivate subjects but, with hindsight, his experiments with Fouraker are also remarkable for the implementation of strict between-subjects anonymity. The latter practice would become very common in later experimental economics, usually as an attempt to implement economic theory's standard atomistic assumptions (especially the ban on other-regarding preferences). Contrary to the standard economic theory, Fouraker and Siegel recognized that interpersonal reactions do matter, but left a systematic investigation of their effects for later research.

More or less simultaneously, Ward Edwards at Michigan pioneered the experimental study of expected utility theory, as axiomatized in the second edition of von Neumann and Morgenstern's *Theory of Games* (1947). Amos Tversky, a student of Edwards and Coombs, would play a major role in the institutionalization of behavioural economics two decades later, as we shall see. In the mid-1950s an interdisciplinary group was also at work on the new theory of individual decision making, under the heading of the Stanford Value Project. Donald Davidson and Pat Suppes (both to become famous later for their contributions to philosophy) published with Siegel one of the first monographs of experimental decision theory (Davidson et al. 1957). At the centre of their research were measurement issues, in particular the implementation of learning theory and Frank Ramsey's method for measuring utilities and subjective probabilities.

Another major centre of interdisciplinary research in those years was the Carnegie group working on the psychology of organizations. Herbert Simon – working at Carnegie and the RAND Corporation, himself a participant in the Santa Monica seminar – is usually credited with being a pivotal player in this connection, although his influence on experimental economics is mostly indirect. The Carnegie group made use of a variety of methodologies, among which experimental 'role playing', 'business games', and simulations were central. In their larger projects, like the Carnegie Tech Management Game, human decision makers took managerial decisions in an environment simulated by a computer. Although primarily devised for pedagogic and illustrative purposes, such games were also used to shed light on the 'boundedly rational' processes of decision making that guide behaviour in big organizations. There is little continuity, however, between this body of work and contemporary experimental economics, with Simon playing a role more as a source of moral support and intellectual inspiration than as a direct contributor to experimental research.

The most famous experimental discovery of this period is due to a scholar who was to have little to do with later developments in experimental economics. Maurice Allais had been developing in France his own version of utility theory as a cardinal measurable quantity well before the publication of the *Theory of Games.* At a conference he organized in Paris in 1952, during a lunch break Allais presented Leonard Savage with a 'questionnaire' that was to become famous as the 'Allais paradox' experiment. When Savage gave answers that were inconsistent with the expected utility model he himself supported, Allais was encouraged to extend his questionnaire and to circulate it more widely.

The results were partially published in French in *Econometrica* (Allais 1953) but received little attention in the short term. The main immediate result of the Allais experiment was Savage's switch to a purely normative defence of expected utility (Jallais and Pradier 2005). Milton Friedman at the time was developing his methodology of positive economics which accorded no importance to the accuracy of the models of individual decision used to predict aggregate phenomena; and Allais's chauvinistic polemic against the 'American School' probably did little to attract sympathy. For about two decades Allais did not pursue research in this area any further.

The only large-scale experimental research project in Europe during this period was led by Reinhard Selten in Frankfurt, under the auspices of Heinz Sauermann. Like other early game theorists, Selten was convinced that the theory could contribute to the solution of important social science problems only if used in conjunction with empirical evidence. Indeed, even his most celebrated theoretical achievement (the concept of subgame perfection) was conceived in the context of a larger experimental project (see Selten 1995).

The last piece of the puzzle of experimental economics in the 1950s is at the same time the most important and the most idiosyncratic. Vernon Smith had been experimenting at Purdue since 1956, focusing on the properties of different market institutions and their effects on the convergence towards equilibrium (see Smith 1981). Smith had an engineering background and, unlike most experimenters at the time, did not approach experiments from a game-theoretic perspective. In the 1940s and 1950s Edward Chamberlin at Harvard had been performing little classroom experiments for illustrative purposes, to show his graduate students the falsity of the competitive theory of markets. Although the results of such experiments had been published in the *Journal of Political Economy* (1948), nobody at the time, including Chamberlin, attributed particular scientific value to them. Smith was the exception: a few years after leaving graduate school he came to question the design used by Chamberlin and to test the robustness of the 'no convergence' results to variations in the exchange institution and repetition of the task.

Overcoming several obstacles, Smith managed to publish his counter-experiments to Chamberlin (Smith 1962). For many years Smith led the only experimental project carried on fully within the boundaries of the economics discipline. In the early 1960s his work received funding from the National Science Foundation, but, apart from a brief attempt to collaborate with the Carnegie

group (see Lee 2004), his work in this phase was mostly carried out in isolation. One important exception is Smith's brief but important encounter with Sidney Siegel at Stanford in 1961. Smith perceived Siegel as much more advanced in methodological matters, and took from him several insights in experimental design that were to become the hallmark of economic experimentation (Smith 1981, 1992).

From the Underground to the Big Bang

Like other innovations of the previous two decades, experimental economics went through a period of slow, quiet growth in the 1960s. Some early contributors, like Allais, disappeared from the scene; others, like Smith, quit experimenting for some time (1967–74) and generally struggled to find an audience. Some areas, like social dilemmas and bargaining experiments, were booming in psychology but had little impact on the economics literature (see Leonard 1994). In the 1970s, however, the landscape of experimental economics changed considerably, partly thanks to the formation of a few key partnerships. During 1968–9 Amos Tversky began collaborating with Daniel Kahneman at the Hebrew University, initially on judgement and then on decision making. In Europe, by 1972, Selten had moved to Bielefeld and started a collaboration with Werner Güth, later author of the first experiments on the ultimatum game. Allais in the meantime returned to expected utility in 1974, and was persuaded to publish a full report in English of his 1952 results (in Allais and Hagen 1979). Allais's legacy would also begin to bear some fruits on the theoretical front. The late 1970s and early 1980s were characterized by a proliferation of alternative models to expected utility, mostly inspired by the experimental evidence that had been accumulated up until then.

After the happy anarchy of the earlier period, the 1970s were marked by the beginning of some controversies and the partial separation of the experimental community into sub-disciplines. In 1974 an article by Tversky and Kahneman in *Science* was widely read as a challenge to the view that human beings were rational agents, and, although it made experiments on judgement and decision making enter the intellectual debate at large, it also fed some deep cross-disciplinary misconceptions. A few years later Lichtenstein and Slovic's seminal experiments on preference reversals were introduced into the economics literature by Grether and Plott (1979), kicking off a series of theoretical and experimental papers that would fill the pages of the *American Economic Review* for years.

Charles Plott had been in close contact with Vernon Smith since the early 1960s, and started to run experiments a decade later, after his move to Caltech. Their collaboration led not only to important experimental projects but also to the creation of the Caltech laboratory and the training of the second and third generations of experimental economists. An important outcome of this period was also the attempt to systematize the methodology of experimental economics around a set of rules or 'precepts' of experimental design (Smith 1976, 1982). Smith in these papers highlighted the importance of monetary incentives to control subjects' preferences, a practice that he had borrowed from Siegel – a psychologist – but that ironically was to become the main distinguishing feature of the 'economic' way of experimenting, as opposed to the more liberal 'psychological' way. With hindsight these methodological papers are also striking for their effective use of the language and conceptual framework of mechanism design theory. In this sense they reflected Smith's (and Plott's) attitude towards the use of experiments to tackle real-world problems of institutional design and policymaking (see Guala 2005).

With the slow exhaustion of general equilibrium theory, the turmoil in macroeconomics, and an increasing disillusionment about econometrics, the 1970s created the conditions for the seeds of the 1940s and 1950s to finally blossom. Experimental economists were in a position to take advantage of this situation. By the early 1980s most of the 'paradigmatic' experiments that would inform subsequent research had already been published (Smith and Plott's experiments on auctions and markets, Lichtenstein and Slovic (1971) on preference reversals, Plott and others on public goods (Isaac et al. 1985), Güth on the

ultimatum game (Güth et al. 1982), Alvin Roth and others on bargaining (Roth and Malouf 1979)). Consolidation meant also differentiation. A persistent low-intensity conflict at the methodological and theoretical level led to the creation of so-called 'behavioural economics'. Whereas experimental economics refers primarily to a method of investigation, the work of behavioural economists is unified by a substantial project of revision of economic theory (especially the replacement of *homo oeconomicus* with a more realistic psychological model), with experimentation constituting a major but by no means exclusive source of evidence.

The history of experimental economics in the 1980s and 1990s is the story of a booming research programme, increasingly influential within the discipline and the social sciences at large, expanding in new directions – neuroscience, for example – and attracting some of the most talented graduate students. Together with game theorists, experimenters have also been increasingly involved in policymaking, notably by contributing to the design of new market institutions for the allocation of sensitive goods – from telecommunication licences to space stations, airport slots, and physicians and surgeons (see Roth 2002). In 2002 the Nobel Memorial Prize in economics awarded to Vernon Smith and Daniel Kahneman provided official acknowledgement of this remarkable revolution.

See Also

- ▶ Allais Paradox
- ▶ Behavioural Game Theory
- ▶ Experimental Economics
- ▶ Field Experiments
- ▶ Kahneman, Daniel (born 1934)
- ▶ Neuroeconomics
- ▶ Smith, Vernon (born 1927)
- ▶ Tversky, Amos (1937–1996)

Bibliography

Allais, M. 1953. La psychologie de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21: 503–546.

Allais, M., and O. Hagen (eds.). 1979. *Expected utility hypothesis and the allais paradox*. Dordrecht: Reidel.

Chamberlin, E.H. 1948. An experimental imperfect market. *Journal of Political Economy* 56: 95–108.

Davidson, D., P. Suppes, and S. Siegel. 1957. *Decision making: An experimental approach*. Stanford: Stanford University Press.

Grether, D., and C. Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69: 623–638.

Guala, F. 2005. *The methodology of experimental economics*. New York: Cambridge University Press.

Güth, W., R. Schmittberger, and B. Schwartz. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3: 363–388.

Isaac, R., K. McCue, and C. Plott. 1985. Public goods provision in an experimental environment. *Journal of Public Economics* 26: 51–74.

Jallais, S., and P. Pradier. 2005. The Allais paradox and its immediate consequences for expected utility theory. In *The experiment in the history of economics*, ed. P. Fontaine and R. Leonard. London: Routledge.

Lee, K. 2004. Rationality, minds, and machines in the laboratory: a thematic history of Vernon Smith's experimental economics. PhD thesis, University of Notre Dame.

Leonard, R. 1994. Laboratory strife: Higgling as experimental science in economics and social psychology. In *Higgling*, ed. N. De Marchi and M. Morgan. Durham, NC: Duke University Press.

Lichtenstein, S., and P. Slovic. 1971. Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89: 46–55.

Mill, J.S. 1836. On the definition of political economy and the method of investigation proper to it. In *Collected works of John Stuart Mill*, vol. 4. Toronto: University of Toronto Press. 1967.

Morgan, M. 2003. Economics. In *The Cambridge history of science, volume 7: The modern social sciences*, ed. T. Porter and D. Ross. Cambridge: Cambridge University Press.

von Neumann, J., and O. Morgenstern. 1947. *The theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press. 1944.

Roth, A. 1995. Introduction to experimental economics. In *The handbook of experimental economics*, ed. J. Kagel and A. Roth. Princeton: Princeton University Press.

Roth, A. 2002. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica* 70: 1341–1378.

Roth, A., and M. Malouf. 1979. Game-theoretic models and the role of information in bargaining. *Psychological Review* 86: 574–594.

Schelling, T. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Selten, R. 1995. Autobiography. In *Les prix Nobel/The Nobel prizes 1994*. Stockholm: Nobel Foundation.

Shubik, M. 1960. Bibliography on simulation, gaming, artificial intelligence and allied topics. *Journal of the American Statistical Association* 55: 736–751.
Siegel, S., and L. Fouraker. 1960. *Bargaining and group decision making*. New York: McGraw-Hill.
Smith, V. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70: 111–137.
Smith, V. 1976. Experimental economics: Induced value theory. *American Economic Review* 66: 274–277.
Smith, V. 1981. Experimental economics at Purdue. In *Papers in experimental economics*. Cambridge: Cambridge University Press.
Smith, V. 1982. Microeconomic systems as an experimental science. *American Economic Review* 72: 923–955.
Smith, V. 1992. Game theory and experimental economics: Beginnings and early influences. In *Towards a history of game theory*, ed. E. Weintraub. Durham, NC: Duke University Press.
Thrall, R., C. Coombs, and R. Davis (eds.). 1954. *Decision processes*. New York: Wiley.
Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1130.

Experimental Labour Economics

Armin Falk and Simon Gächter

Abstract

Experimental labour economics uses experimental techniques to improve our understanding of labour economics issues. We start by putting experimental data into perspective with the data-sets typically used by empirical labour economists. We then discuss several examples of how experiments can inform labour economics.

Keywords

Double auction; Efficiency wages; Employment relation; Experimental labour economics; Gift exchange; Implicit incentives; Incomplete contracts; Involuntary unemployment; Laboratory experiments; Labour economics; Labour market institutions; Minimum wages; Moral hazard; Opportunistic behaviour; Performance incentives; Piece rates; Repeated games; Reservation wage; Tournaments; Wage rigidity

JEL Classifications

C9

Scientific progress relies on testing theories. In labour economics different data sources are available for performing such tests. An important distinction is between circumstantial data and experimental or questionnaire data. Circumstantial data is the by-product of uncontrolled, naturally occurring economic activity. In contrast, experimental data is created explicitly for scientific purposes under controlled conditions. In labour economics, the data most commonly and traditionally used is circumstantial data such as unemployment rates or data on wages, education, or income, complemented by survey data. Labour economists have only recently started to use laboratory experiments.

Laboratory experiments have several important advantages in comparison with data sets typically used in labour economics. A key advantage is the unparalleled opportunity to control crucial aspects of the economic environment. This includes control over information conditions, technology, market structure, and trends in economic fundamentals. Control over the decision environment makes it possible to identify the theoretical equilibrium in an experimental labour market, which is basically impossible with field data. Knowing the equilibrium allows the study of convergence properties, stability and efficiency. Experiments are particularly useful for investigating the economic consequences of important labour market institutions, such as minimum wages or employment protection legislations. The reason is that experiments allow the exogenous changes of institutions, holding everything else constant. In the field, by contrast, institutions are always adopted endogenously. Econometric strategies such as instrumenting for policy changes with political variables can help ameliorate this problem, but do not achieve the unequivocal exogenous variation provided by a laboratory experiment. Laboratory experiments also make it

possible to observe behaviour at the level of individual economic agents. This is important given that theoretical predictions typically involve such micro behaviours. For example, it is possible to directly observe individual reservation wages or individual wage bargaining behaviour. Yet another advantage is that with laboratory experiments one can study, at relatively low cost, institutions that do not yet exist. Analogous to experimental tests of new medicines, where the medication is administered to a small subset of the population initially, laboratory experiments can be used as a first step, before experimenting with institutions in the field. Finally, experimental evidence is replicable, which is a prerequisite in establishing solid empirical knowledge.

Data-Sets and the Comparative Advantage of Laboratory Experiments

Although we believe that laboratory experiments offer important advantages for studying institutions, and should thus be exploited more often, it is important to recognize that there are also drawbacks to this method, which calls for a complementary use of different methods. A potential disadvantage is limited generalizability. Note, however, that this critique holds with respect to any data-set, given that any empirical observation is time and space contingent. Another concern is that experiments may be overly simple, missing potentially relevant aspects of the labour market. This is in fact both a problem and an advantage of experiments. Just as economic models are simpler than reality, so experiments are designed to simplify as much as possible, without losing the essentials. Thus, simplicity need not be a defect of an experiment. The key challenge, just as in the case of building economic models, is to include those features that are essential to the question at hand.

Examples

In this section we discuss a selected set of examples of experiments that were designed to shed light on important issues in labour economics. The examples concern the nature of the employment relationship and its contractual regulation, wage rigidity, performance incentives and their potentially detrimental effects, and labour market institutions.

The Employment Relation

The employment relation is an incomplete contract, which typically leaves many important aspects unspecified. This holds in particular for the content of work effort, which is unregulated and thereby non-enforceable by third parties. Contractual incompleteness gives opportunistic agents an incentive to shirk and therefore leads to an inefficiently low surplus. Thus, voluntary cooperation is necessary to ensure efficiency. Akerlof (1982) argued that many employment relationships are therefore governed by a gift exchange: the firm pays a higher wage than necessary to keep the employee, and the employee returns the gift by providing above minimum effort. Akerlof supported his arguments by a case study and casual observations.

The gift-exchange game by Fehr et al. (1993) provided the first experimental test of the existence of gift exchanges in the framework of a formal game-theoretic model designed to mimic an incomplete employment contract. In their experiment, participants assumed the roles of 'workers' and 'firms'. A firm made a wage offer that a worker could accept. If the worker accepted, he or she had then to choose a costly effort level. Parameters were such that a self-interested worker would always choose the lowest possible effort, since effort was costly. In turn, the firm had no incentive to pay an above-minimal wage, because a self-interested worker would shirk anyway. The results of numerous experiments in this framework showed, however, that wages and effort levels are positively correlated. Higher wages were reciprocated by higher effort levels, a finding which is consistent with the gift-exchange argument by Akerlof. This observation is also consistent with field evidence regarding the link between personnel policy and work morale (Bewley 1999).

In these experiments the employment relationship was modelled as a one-shot game, because this allows an unambiguous prediction under the joint

assumptions of rationality and self-interest. Yet in reality, employment relationships are long-term relationships. To test the impact of repeated interaction, Gächter and Falk (2002) conducted the gift-exchange experiment in the form of repeated games in which the same firm–worker pair interacted for ten periods. These repeated games were compared with one-shot games in which each firm was matched with ten different workers. The results showed a significantly higher effort in the repeated game than in the one-shot games. Gächter and Falk showed that the reason for this result is that in the repeated games the selfish types imitate the reciprocal types. This result provides support for theoretical arguments (for example, MacLeod and Malcomson 1998) that incomplete employment relations allow for implicit incentives for non- opportunistic behaviour.

In the experiments by Gächter and Falk (2002) the experimenter determined the duration of the employment relationship exogenously. In reality, however, the duration of employment relationships arises endogenously. Contract theory suggests that the duration might be linked to contractual incompleteness. Specifically, when contracts are incomplete, a long-term relationship provides implicit incentives that constrain opportunistic behaviour – an argument supported by the cited experimental evidence. If contracts are complete then implicit incentives are not necessary to constrain opportunism. Thus, employment relationships will tend to be short term under contractual completeness. Brown et al. (2004) tested these arguments experimentally and found strong support for them.

Efficiency Wages, Wage Rigidity, and Involuntary Unemployment

Efficiency wage theories explain why even in the absence of market interventions wages might be downwardly rigid, causing involuntary unemployment. Akerlof's (1982) gift-exchange theory is one efficiency wage theory that can explain involuntary unemployment. The main idea is simple. If gift exchanges exist, then firms have no incentive to lower wages because this would lead to low performance. Thus, paying high wages is profitable to the firm – wages are downwardly rigid and can cause involuntary unemployment. Fehr et al. (1998) demonstrated the behavioural validity of this argument experimentally.

Fehr and Falk (1999) provide the most stringent confirmation that gift exchanges can lead to downward wage rigidity. In their experiment an employment relationship was embedded in a 'double auction' market institution in which there were more workers than firms. This institution is known for its competitive properties; under complete contracts experimental double auction markets tend to clear very quickly. In the Fehr–Falk experiments both workers and firms could make wage offers. This enables us to observe whether workers underbid each other and firms therefore have the possibility of employing a worker at a low wage. There was indeed fierce competition among workers who underbid each other down to the theoretically predicted wage. Underbidding occurred in both treatments, the 'complete contract treatment' and the 'incomplete contract treatment'. In the latter, the striking finding was that firms did not take advantage of the possibility of paying low wages; instead they deliberately paid very high wages. The workers' reciprocal effort choice explains why firms had an incentive to pay high wages. In the control experiments with complete contracts gift exchanges were precluded by design and actual wages were very close to market clearing wages. Thus, incomplete contracts and gift exchange can explain wage rigidity and involuntary unemployment.

Performance Incentives (and Their Detrimental Effects)

Compensation and performance incentives have always been central topics in labour economics. Compensation may take different forms. The simplest form is a piece rate where a worker receives a certain wage for each unit she produces. Compensation may also depend on relative performance and be coupled with the possibility of moving up the career ladder. Tournament theory (Lazear and Rosen 1981) is an important theoretical framework for understanding career incentives and relative performance incentives.

Bull et al. (1987) provide the first experimental analysis of piece rate and tournament incentives.

They designed their experiments so that the incentive schemes were directly comparable, that is, the predicted effort level was the same both under piece rates and under tournament incentives. The results confirmed the theoretical predictions in both treatments. As it turned out, however, the support for tournament theory is weaker than for piece rate theory. In various treatment conditions these authors find that average effort choices converged close to the equilibrium prediction, but the variance was up to 30 times higher under tournament incentives than under the piece rate system.

The results by Bull et al. (1987) provide clear evidence that incentives influence behaviour very strongly. However, numerous experiments as well as field evidence (Bewley 1999) suggest that employment relationships are also governed by 'good will' and voluntary cooperation. This raises the question how explicit performance incentives affect voluntary cooperation – a fertile area of current research in experimental labour economics. A nice illustration of the potentially dysfunctional effects of introducing explicit incentives is the field experiment by Gneezy and Rustichini (2000). These authors studied the parents' response to the introduction of a fixed fine for picking up their children too late from kindergarten. The experiment lasted for 20 weeks and there were two conditions. In the baseline condition no fine existed. In the treatment condition the experimenters implemented a fixed fine after week four for picking up a child too late. The fine was removed after week 16. From week seven onwards, there was a steep *increase* in the number of latecomers until their number was roughly twice as high as in the baseline condition. Moreover, when the fine was removed at the end of week 16 the number of tardy parents remained roughly twice as high as in the baseline condition. This result clearly contradicts standard incentive theory, which predicts that the introduction of the fine should lower the incidence of late coming. A likely explanation of this finding is that the implicit contract that governed the employment relationship was changed from a good-will-based one to a market-like transaction, in which 'a fine is a price' and parents bought the commodity of being late.

Labour Market Institutions

A particularly important advantage of laboratory experiments concerns the possibility to test the economic effects of (labour market) institutions in a controlled way. An example of such an institutional test is the paper on minimum wages by Falk et al. (2006). In their experiment firms make wage offers to workers in labour markets either with or without minimum wages. The key insight of their study is that minimum wages may affect the reservation wages of workers in a non-trivial way: first, when minimum wages are introduced, workers stipulate reservation wages above the level of the minimum wage level, because being paid at just the level of the minimum wage is considered unfair. Second, while the introduction of a minimum wage increases reservation wages, the removal of a minimum wage legislation changes reservation wages only marginally. These findings help explain several empirical minimum wage puzzles. First, there exists an anomalously low utilization of subminimum wages in situations where employers actually could pay workers less than the minimum; second there exist so-called spillover effects, that is, wages are often increased by an amount in excess of that necessary for compliance with the minimum wage; and third, minimum wages do not always cause a decrease in employment, in particular if the minimum wage increase is modest (see also Card and Krueger 1995).

The finding that minimum wages affect workers' fairness perceptions of wages is also supported by Brandts and Charness (2004) who introduced a minimum wage in the context of an experimental labour market with worker moral hazard where workers' fairness concerns drive effort. They show that workers provide less effort for the same wage level in the presence of the minimum wage. This supports the view that the impact of minimum wages on workers' attributions of fairness intentions to firms partially shapes their effort responses.

Concluding Remarks

Experimental economics is a method of empirical investigation, not a separate subfield of

economics. Experimental methods can therefore in principle be utilized in all areas of economics. In this article we have illustrated some selected applications of experimental methods to important issues in labour economics. Further discussions of the issues raised here can be found in Fehr and Gächter (2000), Gächter and Fehr (2002), Fehr and Falk (2002), Falk and Fehr (2003) and Falk and Huffman (2007).

See Also

- ▶ Behavioural Economics and Game Theory
- ▶ Incomplete Contracts
- ▶ Institutional Economics
- ▶ Labour Economics
- ▶ Personnel Economics
- ▶ Reciprocity and Collective Action

Bibliography

Akerlof, G. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97: 543–569.

Bewley, T. 1999. *Why wages don't fall in a recession*. Cambridge, MA: Harvard University Press.

Brandts, J., and G. Charness. 2004. Do labour market conditions affect gift exchange? Some experimental evidence. *Economic Journal* 114: 684–708.

Brown, M., A. Falk, and E. Fehr. 2004. Relational contracts and the nature of market interactions. *Econometrica* 72: 747–780.

Bull, C., A. Schotter, and K. Weigelt. 1987. Tournaments and piece rates: An experimental study. *Journal of Political Economy* 95: 1–33.

Card, D., and A. Krueger. 1995. *Myth and measurement: The new economics of the minimum wage*. Princeton: Princeton University Press.

Falk, A., and E. Fehr. 2003. Why labor market experiments? *Labor Economics* 10: 399–406.

Falk, A., and D. Huffman. 2007. Studying labor market institutions in the lab: Minimum wages, employment protection and workfare. *Journal of Institutional and Theoretical Economics* 163: 30–45.

Falk, A., E. Fehr, and C. Zehnder. 2006. Fairness perceptions and reservation wages: The behavioral effects of minimum wage laws. *Quarterly Journal of Economics* 121: 1347–1381.

Fehr, E., and A. Falk. 1999. Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy* 107: 106–134.

Fehr, E., and A. Falk. 2002. The psychological foundations of incentives. *European Economic Review* 46: 687–724.

Fehr, E., and S. Gächter. 2000. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14(3): 159–181.

Fehr, E., G. Kirchsteiger, and A. Riedl. 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108: 437–460.

Fehr, E., E. Kirchler, A. Weichbold, and S. Gächter. 1998. When social norms overpower competition – Gift exchange in experimental labor markets. *Journal of Labor Economics* 16: 324–351.

Gächter, S., and A. Falk. 2002. Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics* 104: 1–27.

Gächter, S., and E. Fehr. 2002. Fairness in the labour market: A survey of experimental results. In *Surveys in experimental economics. Bargaining, cooperation and election stock markets*, ed. B. Friedel and M. Lehmann- Waffenschmidt. Heidelberg/New York: Physica.

Gneezy, U., and A. Rustichini. 2000. A fine is a price. *Journal of Legal Studies* 29: 1–17.

Lazear, P., and S. Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.

MacLeod, B., and J. Malcomson. 1998. Motivation and markets. *American Economic Review* 88: 388–411.

Experimental Macroeconomics

John Duffy

Abstract

Experimental macroeconomics is a sub-field of experimental economics that makes use of controlled laboratory methods to understand aggregate economic phenomena and to test the specific assumptions and predictions of macroeconomic models. This article reviews important contributions of experimental macroeconomics research, which include an understanding of when equilibration works, when it fails, and the means by which macro-coordination problems are resolved. It also discusses important methodological issues including the choice of market institution, the implementation of representative agent and overlapping generations models, discounting and infinite horizons, and the external validity of experimental macroeconomic findings.

Keywords

Anchoring effects; Asset pricing; Bubbles; Business cycles; Contagion; Coordination failure; Double auction; Equilibration; Equilibrium selection; Experimental macroeconomics; Forecasting; General equilibrium; Infinite horizons; Laffer curve; Learning; Microfoundations; Money-search models; Multiple equilibria; Optimal growth; Overlapping generations models; Partial equilibrium; Representative agent; Search models of money; Speculative attacks; Strategic uncertainty; Sunspot variables; Time consistency

JEL Classifications

C9

Experimental macroeconomics is a subfield of experimental economics that makes use of controlled laboratory methods to understand aggregate economic phenomena and to test the specific assumptions and predictions of macroeconomic models. Surveys of experimental macroeconomics are found in Ochs (1995), Duffy (1998) and Ricciuti (2004). Macroeconomic topics that have been studied in the laboratory include convergence to Walrasian competitive equilibrium (Lian and Plott 1998), growth and development (Lei and Noussair 2002; Capra et al. 2005), specialization and trade (Noussair et al. 1995), Keynesian coordination failures (Cooper 1999; Van Huyck et al. 1990), the use of money as a medium of exchange (Brown 1996; Duffy and Ochs 1999, 2002) and as a store of value (McCabe 1989; Lim et al. 1994; Marimon and Sunder 1993, 1994), exchange rate determination (Arifovic 1996; Noussair et al. 1997), money illusion (Fehr and Tyran 2001), asset price bubbles and crashes (Smith et al. 1988; Lei et al. 2001; Hommes et al. 2005) sunspots (Marimon et al. 1993; Duffy and Fisher 2005), bank runs (Schotter and Yorulmazer 2003; Garratt and Keister 2005), contagions (Corbae and Duffy 2006), speculative currency attacks (Heinemann et al. 2004), and the economic impact of various fiscal and monetary policies (Riedl and Van Winden 2001; Arifovic and Sargent 2003; Marimon and Sunder 1994; Bernasconi and Kirchkamp 2000).

The use of laboratory experiments, involving small groups of subjects interacting with one another for short periods of time, to analyse aggregate, economy-wide phenomena or to test macroeconomic model predictions or assumptions might be met with some scepticism. However, there are many insights to be gained from controlled laboratory experimentation that cannot be obtained using standard macroeconometric approaches, namely, econometric analyses of the macroeconomic data reported by government agencies. Often the data most relevant to testing a macroeconomic model are simply unavailable. There may also be identification, endogeneity and equilibrium selection issues that cannot be satisfactorily addressed using econometric methods. Indeed, Robert Lucas (1986) was the first macroeconomist to make such observations, and he invited laboratory tests of rational expectations macroeconomic models; much of the subsequent experimental macroeconomics literature may be viewed as a response to Lucas's (1986) invitation. It is also worth noting that experimental methodologies have been improbably applied to the study of many other aggregate phenomena including astronomy, epidemiology, evolution, meteorology, political science and sociology.

Insights from Macroeconomic Experiments

To date, experimental macroeconomics research has yielded some important insights, including an understanding of when equilibration works, when it fails, and the means by which equilibrium selection or coordination problems are resolved. Equilibration, the process by which competitive equilibrium is achieved, is often ignored by modern macroeconomic modellers, who typically assume that market clearing is friction-free and instantaneous. Experimentalists, following the lead of Smith (1962), have explored mechanisms such as the double auction, the availability of information, futures markets and other means by which this equilibration

might be achieved or enhanced (see, for example, Forsythe et al. 1982; Plott and Sunder 1982; Sunder 1995 for partial equilibrium approaches; and Lian and Plott 1998 for a general equilibrium approach). A general finding is that, with enough trading experience and information feedback about transaction prices, bids, and asks, even small populations of five to ten subjects can learn to trade at prices and achieve efficiency consistent with competitive equilibrium in a large class of market environments. Indeed, the institutional rules, for instance of the double auction, may be all that is necessary to assure equilibration, as shown in the zero-intelligence trader approach of Gode and Sunder (1993).

Experimental insights regarding equilibration have enabled experimentalists to design market environments where equilibration may fail to obtain; in its place are observed price bubbles and crashes (Smith et al. 1988; Lei et al. 2001; Hommes et al. 2005). Explaining these laboratory asset price bubbles has proved challenging. Lei et al. (2001) show that speculative motives alone cannot explain bubble formation and suggest that it may have more to do with subject boredom. Duffy and Ünver (2006) suggest that anchoring effects may factor in subjects' bidding up of prices until binding budget constraints force a crash. A further puzzle is that experienced subjects in laboratory asset markets learn to avoid price bubbles and crashes, and generally price assets in line with fundamental values. An explanation for why bubbles and crashes occur among inexperienced but not experienced subjects has yet to be provided. Experiments with mixtures of experienced and inexperienced subjects show no tendency for bubbles to arise (Dufwenberg et al. 2005).

In environments with multiple equilibria, theory is typically silent as to which equilibrium agents will select or whether there will be transitions between equilibria. Understanding how agents coordinate on an equilibrium is of great interest to macroeconomists, as coordination problems are thought to play an important role in the persistence of business cycle fluctuations. Experimental evidence can and has been used to address the issue of which, among multiple equilibria, is most likely to be achieved, and why.

For instance, Van Huyck et al. (1990) have shown how minimum effort, team production payoff functions can lead to Keynes-type coordination failures – that is, coordination by groups of subjects on Pareto inferior equilibria. Such inefficiencies do not arise from conflicting objectives or from asymmetries of information; rather, they arise from individuals' *strategic uncertainty* with regard to the actions of other market participants. Similarly, Duffy and Ochs (1999, 2002) report that subjects have no difficulty coordinating on efficient monetary exchange equilibria in Kiyotaki–Wright-type money-search models when theory calls for the use of fundamental, cost-minimizing strategies, but subjects have much greater difficulty coordinating on efficient monetary equilibria that require them to employ more costly and forward-looking, speculative strategies, due perhaps to the unwillingness of other subjects to adopt those same speculative strategies.

Not all the experimental evidence points to inefficiencies in macro-coordination problems. Marimon and Sunder (1993) show that when subjects are presented with a Laffer-curve-type tradeoff between two inflation rates, the efficient, low-inflation equilibria is more likely to be selected than is the inefficient, high-inflation equilibrium. They show that the low-inflation equilibrium is stable under the adaptive learning dynamics that subjects use whereas the high-inflation equilibrium is not. Similarly, Arifovic and Sargent (2003) study behaviour in a Kydland–Prescott model of expected inflation output trade-offs and find that a majority of subjects acting in the role of central bank are able to choose policies so as to induce subjects, in the role of the public, to coordinate their expectations on the efficient but time- inconsistent Ramsey equilibrium. Still, they report occasional instances of 'backsliding' to the less efficient, time-consistent Nash equilibrium.

Finally, Duffy and Fisher (2005) explore subjects' use of non-fundamental 'sunspot' variables as coordination devices in an environment with multiple equilibria. They show that, when information is highly centralized, as in a call market, subjects use realizations of a sunspot variable as a device for coordinating on low- or high-price equilibria, but that this coordination mechanism

may break down when information is more decentralized, as in a double auction, or when the mapping from realizations of the sunspot variable to the action space is unclear.

Methodological Issues

Methodologically, macroeconomic experiments typically involve some kind of centralized market-clearing mechanism through which subjects interact with one another, for instance as buyers or sellers, or both. The double auction market mechanism (Friedman and Rust 1991) is the most commonly used market-clearing mechanism, as it allows for continuous information on bids, asks, transaction prices and volume – information which is thought to be critical to rapid equilibration and high levels of allocative efficiency (Lian and Plott 1998; Noussair et al. 1995, 1997). The simultaneous, sealed-bid 'call' market version of this mechanism has also been used by some researchers (Cason and Friedman 1997; Duffy and Fisher 2005; Capra et al. 2005).

Some less centralized market mechanisms have also been used. For instance, Brown (1996), Duffy and Ochs (1999, 2002) study a money-search model in which subjects are randomly paired and may trade goods with one another at a fixed exchange rate. In addition, game-theoretic models are also commonly employed, especially in studies of coordination failure, contagion and speculative attacks (Van Huyck et al. 1990; Corbae and Duffy 2006; Heinemann et al. 2004).

A hallmark of modern macroeconomic modelling is the characterization of the economy using recursive dynamical systems where expectations of future endogenous variables determine current outcomes. Several experimental researchers testing such models have found it useful to separate subjects' forecast decisions from market-trading decisions. For instance, Marimon and Sunder (1993, 1994, 1995) and Hommes et al. (2005) elicit subjects' forecasts of the next period's price level. Using these individual forecasts, they determine subjects' individual demands for the consumption good in the current period and, as supply is fixed, they simultaneously determine the current period price. Similarly, Adam (2007) elicits forecasts of inflation one and two periods ahead, consistent with the monetary sticky price model that he investigates; these expectations are then used to determine output and inflation in the current period. Marimon and Sunder (1994) refer to this type of experimental design as a 'learning to forecast' framework, which they contrast with a 'learning to optimize' framework. Of course, in macroeconomic models, it is assumed that agents are able to both forecast and optimize at the same time.

Many macroeconomic models have representative agents and infinite horizons or an infinity of agents and goods which pose some challenges for laboratory implementation and testing of theoretical predictions. The representative agent assumption has been examined by Noussair and Matheny (2000) and Lei and Noussair (2002). They compare consumption and investment decisions made by individual subjects operating as 'representative agent-social planners' in the standard Cass–Koopmans optimal growth framework with the decisions made by groups of subjects who first trade shares of capital via a double-auction market clearing mechanism and then allocate their income between consumption and investment. They find that the double-auction market mechanism results in allocations that are far closer to the theoretical predictions than are the decisions made by subjects in the representative agent role attempting to solve the optimization problem on their own.

To implement infinite horizons, researchers have adopted two designs. One design, used for example by Marimon and Sunder (1993), is to recruit subjects for a fixed period of time but terminate the session early, without advance notice, following the end of some period of play. As Marimon and Sunder use a forward- looking dynamic model, they use the one-step-ahead forecasts made by a subset of subjects who are paid for their forecast accuracy to determine final period allocations. A second design is to introduce a constant small probability, $1 - \delta$, that each period will be the last one played in a sequence, and allow enough time for several indefinite sequences to be played in an experimental session (Duffy and Ochs

1999, 2002; Lei and Noussair 2002; Capra et al. 2005). This design has the advantage of inducing both the stationarity associated with an infinite horizon and discounting of future payoffs at rate $(1 - \delta)/\delta$ per period (equivalently a discount *factor* of δ). Related to the infinite horizon problem, overlapping generations models, as studied by Marimon and Sunder (1993, 1994, 1995) and Marimon et al. (1993) have an infinity of agents (and goods). Marimon and Sunder cope with this difficulty by recycling subjects – allowing each subject to live several two- period lives over the course of an indefinite sequence of periods. Marimon and Sunder (1993) argue that this repeated entry and exit of subjects does not induce any strategic opportunities that are not already present in the overlapping generations model without 'rebirth'. Indeed, the need for a large number of agents to study macroeconomic behaviour is a common issue confronted by researchers. However, results from many double auction experiments suggest that competitive equilibrium can be quickly achieved with as few as three to five subjects operating on each side of the market. Similarly, while search models of money assume a continuum of agents, Duffy and Ochs (2002) argue that the strategic incentives generated by having finite subject populations do not alter the equilibrium predictions of those models under the assumption of a continuum of agents.

Perhaps the most difficult methodological issue is the external validity of macroeconomic experimental findings. While external validity is generally a problem for all experimental economists, it might be regarded as a greater problem for macroexperimentalists seeking to explain economy-wide aggregate macroeconomic phenomena using necessarily small-scale laboratory evidence. Experimental macroeconomists have several responses to this issue. First, as noted earlier, modern macroeconomic models have explicit microfoundations as to how individual agents make decisions (for example, agents recognize the relevant trade-offs, form rational expectations) which can be directly tested in the laboratory. Indeed, in the laboratory one can be more certain about micro-level *causal relationships*, that is, that an experimenter induced change in a variable is the source of any observed change in subject behaviour as opposed to some other, unaccounted-for factors. Macroeconometric analyses of field data cannot claim the same degree of *internal validity*. A second response is to make use of highly experienced subjects – those who have participated in the same experiment many times – as a means of better proxying real-world behaviour. As noted earlier, asset price bubbles and crashes seem to disappear with experienced subjects. A third response has been to use parametric forms or calibrations of macroeconomic models that are of interest to macroeconomists, or to present subjects with real macroeconomic data as part of the experimental design (for example, Bernasconi et al. 2004). Finally, many experimentalists would argue that all experimental work, including macroeconomic experiments, should be judged by the findings obtained and not by biases concerning the suitability of laboratory versus other empirical methods, all of which have their strengths and weaknesses.

See Also

- Behavioural Finance
- Coordination Problems and Communication
- Experimental Economics

Bibliography

Adam, K. 2007. Experimental evidence on the persistence of output and inflation. *Economic Journal* 117: 603–636.

Arifovic, J. 1996. The behavior of the exchange rate in the genetic algorithm and experimental economies. *Journal of Political Economy* 104: 510–541.

Arifovic, J., and T.J. Sargent. 2003. Laboratory experiments with an expectational Phillips curve. In *Evolution and procedures in central banking*, ed. D.E. Altig and B.D. Smith. Cambridge: Cambridge University Press.

Bernasconi, M., and O. Kirchkamp. 2000. Why do monetary policies matter? An experimental study of saving and inflation in an overlapping generations model. *Journal of Monetary Economics* 46: 315–343.

Bernasconi, M., O. Kirchkamp, and P. Paruolo. 2004. *Do fiscal variables affect fiscal expectations? Experiments with real world and lab data*. SPF 504 Discussion Paper No. 04–26. University of Mannheim.

Brown, P.M. 1996. Experimental evidence on money as a medium of exchange. *Journal of Economic Dynamics and Control* 20: 583–600.

Capra, C.M., T. Tanaka, C.F. Camerer, L. Munyan, V. Sovero, L. Wang, and C. Noussair. 2005. The impact of simple institutions in experimental economies with poverty traps. Working paper.

Cason, T.N., and D. Friedman. 1997. Price formation in single call markets. *Econometrica* 65: 311–345.

Cooper, R.W. 1999. *Coordination games*. Cambridge: Cambridge University Press.

Corbae, D., and J. Duffy. 2006. Experiments with network formation. Working paper.

Duffy, J. 1998. Monetary theory in the laboratory. *Federal Reserve Bank of St. Louis Economic Review* 80: 9–26.

Duffy, J., and E.O.N. Fisher. 2005. Sunspots in the laboratory. *American Economic Review* 95: 510–529.

Duffy, J., and J. Ochs. 1999. Emergence of money as a medium of exchange: An experimental study. *American Economic Review* 89: 847–877.

Duffy, J., and J. Ochs. 2002. Intrinsically worthless objects as media of exchange: Experimental evidence. *International Economic Review* 43: 637–673.

Duffy, J., and U. Ünver. 2006. Asset price bubbles and crashes with near-zero-intelligence traders. *Economic Theory* 27: 537–563.

Dufwenberg, M., T. Lindqviist, and E. Moore. 2005. Bubbles and experience: An experiment. *American Economic Review* 95: 1731–1737.

Fehr, E., and J.-F. Tyran. 2001. Does money illusion matter? *American Economic Review* 91: 1239–1262.

Forsythe, R., T.R. Palfrey, and C.R. Plott. 1982. Asset valuation in an experimental market. *Econometrica* 58: 537–568.

Friedman, D., and J. Rust. 1991. *The double auction market: Institutions, theories and evidence*. Cambridge, MA: Perseus Publishing.

Garratt, R., and T. Keister. 2005. Bank runs: An experimental study. Working paper, University of California, Santa Barbara.

Gode, D.K., and S. Sunder. 1993. Allocative efficiency of markets with zero intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy* 101: 119–137.

Heinemann, F., R. Nagel, and P. Ockenfels. 2004. The theory of global games on test: Experimental analysis of coordination games with public and private information. *Econometrica* 72: 1583–1599.

Hommes, C.H., J. Sonnemans, J. Tuinstra, and H. van de Velden. 2005. Coordination of expectations in asset pricing experiments. *Review of Financial Studies* 18: 955–980.

Lei, V., and C.N. Noussair. 2002. An experimental test of an optimal growth model. *American Economic Review* 92: 549–570.

Lei, V., C.N. Noussair, and C.R. Plott. 2001. Nonspeculative bubbles in experimental asset markets: Lack of common knowledge of rationality vs. actual irrationality. *Econometrica* 69: 831–859.

Lian, P., and C.R. Plott. 1998. General equilibrium, markets, macroeconomics and money in a laboratory experimental environment. *Economic Theory* 12: 21–75.

Lim, S.S., E.C. Prescott, and S. Sunder. 1994. Stationary solution to the overlapping generations model of fiat money: Experimental evidence. *Empirical Economics* 19: 255–277.

Lucas, R.E. 1986. Adaptive behavior and economic theory. *Journal of Business* 59: S401–S426.

Marimon, R., S.E. Spear, and S. Sunder. 1993. Expectationally driven market volatility: An experimental study. *Journal of Economic Theory* 61: 74–103.

Marimon, R., and S. Sunder. 1993. Indeterminacy of equilibria in a hyperinflationary world: Experimental evidence. *Econometrica* 61: 1073–1107.

Marimon, R., and S. Sunder. 1994. Expectations and learning under alternative monetary regimes: An experimental approach. *Economic Theory* 4: 131–162.

Marimon, R., and S. Sunder. 1995. Does a constant money growth rule help stabilize inflation? Experimental evidence. *Carnegie-Rochester Conference Series on Public Policy* 43: 111–156.

McCabe, K.A. 1989. Fiat money as a store of value in an experimental market. *Journal of Economic Behavior and Organization* 12: 215–231.

Noussair, C.N., and K.J. Matheny. 2000. An experimental study of decisions in dynamic optimization problems. *Economic Theory* 15: 389–419.

Noussair, C.N., C.R. Plott, and R.G. Riezman. 1995. An experimental investigation of the patterns of international trade. *American Economic Review* 85: 462–491.

Noussair, C.N., C.R. Plott, and R.G. Riezman. 1997. The principles of exchange rate determination in an international financial experiment. *Journal of Political Economy* 105: 822–861.

Ochs, J. 1995. Coordination problems. In *The handbook of experimental economics*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.

Plott, C.R., and S. Sunder. 1982. Efficiency of experimental security markets with insider information: An application of rational-expectations models. *Journal of Political Economy* 90: 663–698.

Ricciuti, R. 2004. Bringing macroeconomics into the lab. Working paper No. 26, International Centre for Economic Research.

Riedl, A., and F. van Winden. 2001. Does the wage tax system cause budget deficits? A macro-economic experiment. *Public Choice* 109: 371–394.

Schotter, A., and T. Yorulmazer. 2003. On the severity of bank runs: An experimental study. Working paper, Center for Experimental Social Science, New York University.

Smith, V.L. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70: 111–137.

Smith, V.L., G.L. Suchanek, and A.W. Williams. 1988. Bubbles, crashes, and endogenous expectations in

experimental spot asset markets. *Econometrica* 56: 1119–1151.
Sunder, S. 1995. Experimental asset markets: A survey. In *The handbook of experimental economics*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.
Van Huyck, J.B., R.C. Battalio, and R.O. Beil. 1990. Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review* 80: 234–248.

Experimental Methods in Economics

C. F. Bastable

Abstract

In the mid-20th century economists became involved in the design and conduct of laboratory experiments to examine propositions implied by economic theory. This development brought new standards of rigour to the data gathering process. This article gives an account of the author's experiment in 1956 to test the hypothesis that the competitive market process yields welfare improving (and, under certain limiting ideal conditions, welfare maximizing) outcomes, provides an interpretive history of the development of experimental economics, discusses the functions of market experiments in microeconomic analysis, and classifies the application of experimental methods.

Keywords

Agenda processes; Antitrust economics; Auctions (theory); Buyer's surplus; Collusion; Compensated unanimity processes; Competitive equilibrium; Contestable markets hypothesis; Decentralization; Demand Functions; Dominance; Double oral auction; Expected utility theory; Experimental methods in economics; Hypotheses; Incentive compatibility; Industrial organization; Institutions; Labour supply; Market as positive sum game; Monopoly; Monopsony; Natural monopoly; Nonsatiation; Parallelism; Preference theory; Privacy; Public enforcement of law; Public good allocation; Research programmes; Saliency; Scale economics; seller's surplus; Smith, A.; Smith, V. L.; Speculative gains; Supply functions; Testing; Vickrey, W.; Vouchers

JEL Classifications

C9

Historically, the method and subject matter of economics have presupposed that it was a non-experimental (or 'field observational') science more like astronomy or meteorology than physics or chemistry. Based on general, introspectively 'plausible', assumptions about human preferences, and about the cost and technology based supply response of producers, economists have sought to understand the functioning of economies, using observations generated by economic outcomes realized over time. The data of the astronomer is of this same type, but it would be wrong to conclude that astronomy and economics are methodologically equivalent. There are two important differences between astronomy and economics which help to illuminate some of the methodological problems of economics. First, based upon parallelism (the maintained hypothesis that the same physical laws hold everywhere), astronomy draws on all the relevant theory from classical mechanics and particle physics – theory which has evolved under rigorous laboratory tests. Traditionally, economists have not had an analogous body of tested behavioural principles that have survived controlled experimental tests, and which can be assumed to apply with insignificant error to the microeconomic behaviour that underpins the observable operations of the economy. Analogously, one might have supposed that there would have arisen an important area of common interest between economics and, say, experimental psychology, similar to that between astronomy and physics, but this has only started to develop in recent years.

Second, the data of astronomy are painstakingly gathered by professional observational astronomers for scientific purposes, and these data are taken seriously (if not always non-controversially) by astrophysicists and

cosmologists. Most of the data of economics has been collected by government or private agencies for non- scientific purposes. Hence astronomers are directly responsible for the scientific credibility of their data in a way that economists have not been. In economics, when things appear not to turn out as expected the quality of the data is more likely to be questioned than the relevance and quality of the abstract reasoning. Old theories fade away, not from the weight of falsifying evidence that catalyses theoretical creativity into developing better theory, but from lack of interest, as intellectual energy is attracted to the development of new techniques and to the solution of new puzzles that remain untested.

At approximately the mid-20th century, professional economics began to change with the introduction of the laboratory experiment into economic method. In this embryonic research programme economists (and a psychologist, Sidney Siegel) became directly involved in the design and conduct of experiments to examine propositions implied by economic theories of markets. For the first time this made it possible to introduce *demonstrable* knowledge into the economist's attempt to understand markets.

This laboratory approach to economics also brought to the economist direct responsibility for an important source of scientific data generated by controlled processes that can be replicated by other experimentalists. This development invited economic theorists to submit to a new discipline, but also brought an important new discipline and new standards of rigour to the data gathering process itself.

An untested theory is simply a hypothesis. As such it is part of our *self*- knowledge. Science seeks to expand our knowledge of *things* by a process of testing this type of self-knowledge. Much of economic theory can be called, appropriately, 'ecclesiastical theory'; it is accepted (or rejected) on the basis of authority, tradition, or opinion about assumptions, rather than on the basis of having survived a rigorous falsification process that can be replicated.

Interest in the replicability of scientific research stems from a desire to answer the question 'Do you see what I see?' Replication and control are the two primary means by which we attempt to reduce the error in our common knowledge of economic processes. However, the question 'Do you see what I see?' contains three component questions, recognition of which helps to identify three different senses in which a research study may fail to be replicable:

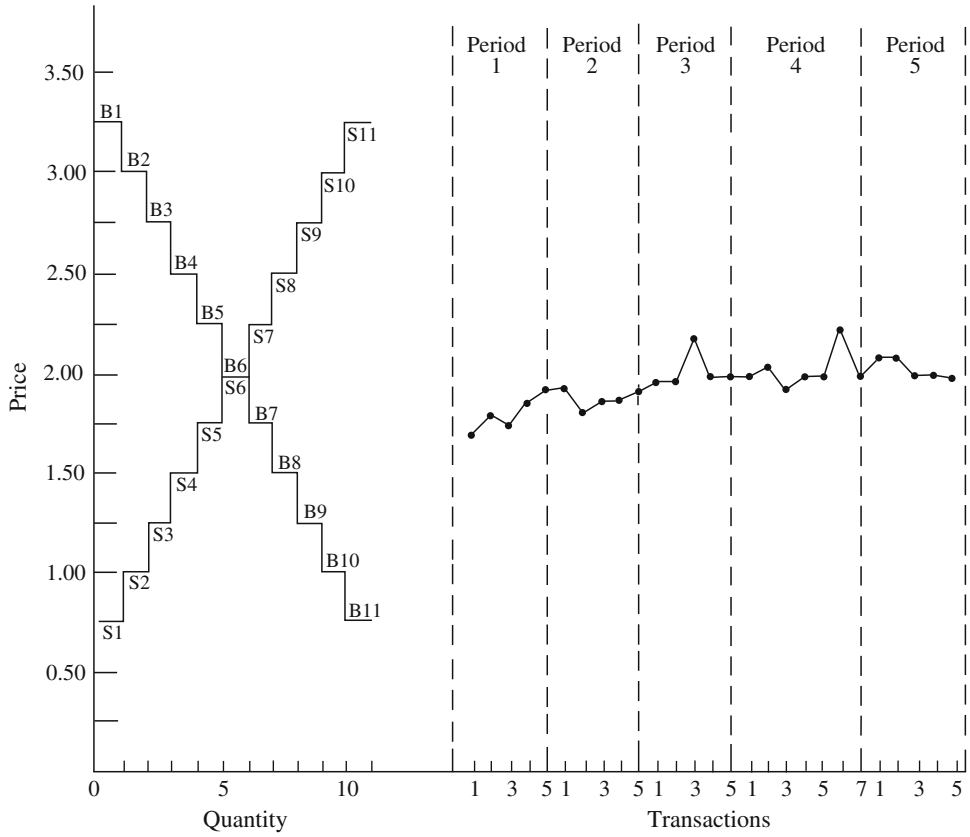
(1) *Do you observe what I observe?* Since economics has traditionally been confined to the analysis of non-experimental data, the answer to this question has been trivially, 'yes'. We observe the same thing because we use the same data. This non-replicability of our traditional data sources has helped to motivate some to turn increasingly to experimental methods. We can say that you have replicated my experiments if you are unable to reject the hypothesis that your experimental data came from the same population as mine. This means that the experimenter, his/her subjects, and/or procedures are not significant treatment variables.
(2) *Do you interpret what we observe as I interpret it?* Given that we both observe the same, or replicable data, do we put the same interpretation on these data? The interpretation of observations requires theory (either formal or informal), or at least an empirical interpretation of the theory in the context that generated the data. Theory usually requires empirical interpretation either because (i) the theory is not developed directly in terms of what can be observed (e.g. the theory may assume risk aversion which is not directly observable), or (ii) the data were not collected for the purpose of testing, or estimating the parameters of a theory. Consequently, failure to replicate may be due to differences in interpretation which result from different meanings being ascribed to the theory. Thus two researchers may apply different transformations to raw field data (e.g. different adjustments for the effect of taxes), so that the results are not replicable because their theory interpretations differ.
(3) *Do you conclude what I conclude from our interpretation?* The conclusions reached in two different research studies may be different

even though the data and their interpretation are the same. In economics this is most often due to different model specifications. This problem is inherent in non-experimental methodologies in which, at best, one usually can estimate only the parameters of a pre-specified model and cannot credibly test one model or theory against another. An example is the question of whether the Phillips' curve constitutes a behavioural trade-off between the rates of inflation and unemployment, or represents an equilibrium association without causal significance.

Markets and Market Experiments

Markets and how they function constitute the core of any economic system, whether it is highly decentralized – popularly, a 'capitalistic' system, or highly centralized – popularly, a 'planned' system. This is true for the decentralized economy because markets are the spontaneous institutions of exchange that use prices to guide resource allocation and human economic action. It is true for the centralized economy because in such economies markets always exist or arise in legal form (private agriculture in Russia) and clandestine or illegal form (barter, bribery, the trading of favours, and underground exchange in Russia, Poland and elsewhere). Markets arise spontaneously in all cultures in response to the human desire for betterment (to 'profit') through exchange. Where the commodity or service is illegal (prostitution, gambling, the sale of liquor under Prohibition or of marijuana, cocaine, etc.) the result is not to prevent exchange, but to raise the risk and therefore the costs of exchange. This is because enforcement is itself costly, and it is never economical for the authorities (whether Soviet or American) even to approximate perfect enforcement. The spontaneity with which markets arise is perhaps no better illustrated than when (1979–80) US airlines for promotional purposes issued travel vouchers to their passengers. One of these vouchers could be redeemed by the bearer as a cash substitute in the purchase of new airline tickets. Consequently vouchers were of value to future passengers. Furthermore, since (as Hayek would say) the 'circumstances of time and place' for the potential redemption of vouchers were different for different individuals, there existed the preconditions for the active voucher market that was soon observed in all busy airports. Current passengers with vouchers who were unlikely to be travelling again soon held an asset worth less to themselves than to others who were more certain of their future or impending travel plans. The resulting market established prices that were discounts from the redemption or 'face' value of vouchers. Sellers who were unlikely to be able to redeem their vouchers preferred to sell them at a discount for cash. Buyers who were reasonably sure of their travel plans could save money by purchasing vouchers at a discount. Thus the welfare of every active buyer and seller increased via this market. Without a market, many – perhaps most – vouchers would not have been exercised and would thus have been 'wasted'.

The previous paragraph illustrates a fundamental hypothesis (theorem) of economics: the ('competitive') market process yields welfare improving (and, under certain limiting ideal conditions, welfare maximizing) outcomes. But is the hypothesis 'true', or at least very probably true? (Lakatos (1978) would correctly ask 'Has it led to an empirically progressive research programme?') I think it is 'true', but how do I know this? Do you see what I see? A Marxist does not see what I see in the above interpretation of a market. The young student studying economics does not see what I see, although if they continue to study economics eventually they (predictably) come to see what I see (or, at least, they say they do). Is this because we have inadvertently brainwashed them? The gasoline consumer does not see what I see. They see themselves in a *zero* sum game with an oil company: any increase in price merely redistributes wealth from the consumer to the company, which is not 'fair' since the company is richer. What I see in a market is a *positive* sum game yielding gains from exchange, which constitutes the fundamental mechanism for creating, not merely redistributing wealth. The traditional method by which the economist gets others to see this 'true' function of markets is by



Experimental Methods in Economics, Fig. 1

logical arguments (suppose it were not true, then . . .), examples, and 'observations', such as are contained in my description of the voucher market, in which what is 'observed' is hortatively described and interpreted in terms of the hypothesis itself. But if this knowledge of the function of markets is 'true', can it be demonstrated? Experimentalists claim that laboratory experiments can provide a uniquely important technique of demonstration for supplementing the theoretical interpretation of field observations.

I conducted my first experiment in the spring of 1956. Since then hundreds of similar, as well as environmentally richer experiments have been conducted by myself and by others. In 1956, my introductory economics class consisted of 22 science and engineering students, and although this might not have been the 'large number' traditionally thought to have been necessary to yield a competitive market, I thought it was large enough for a practice run to initiate a research programme capable of falsifying the standard theory. I conducted the experiment before lecturing on the theory and 'behaviour' of markets in class so as not to 'contaminate' the sample. The 22 subjects were each assigned one card from a well-shuffled deck of 11 white and 11 yellow cards. The white cards identified the sellers, and the yellow cards identified the buyers. Each white card carried a price, known only to that seller, which represented that seller's minimum selling price for one unit, and each yellow card identified a price, known only to that buyer, representing that buyer's maximum buying price for one unit. On the left of Fig. 1 is listed these so-called 'limit' prices, identified by buyer, B1, B2 etc. (in descending order, D) and by seller, S1, S2 etc. (in ascending order, S). To keep things simple and well controlled each buyer (seller) was informed that he/she was a buyer (seller) of at most one unit of

the item in each of several trading periods. Thus demand, D (supply, S) was 'renewed' in each trading period as a steady state flow, with no carry-over in unsatisfied demand (or unsold stock), from one period to the next. In the airline voucher example, imagine the vouchers being issued, followed by trading; the vouchers then expire, new vouchers are issued, traded and so on. In the experiment, suppose real motivation is provided by promising to pay (in cash) to each buyer the difference between that buyer's assigned limit buying price and the price actually paid in each period that a unit is purchased in the market. Thus suppose seller 5 sells their unit to buyer 2 at the price 2.25. Then buyer 2 earns a 'profit' of $0.75 from this exchange. In this way we induce on each buyer a value (or hypothesized willingness-to-pay) equal to the assigned limit buy price. Similarly, suppose each seller is paid the difference between that seller's actual sales price and assigned limit price ('cost', or willingness-to-sell) in each trading period that a unit is sold. Thus in the previous exchange example, seller 5 earns $0.50 from the transaction.

This experimental procedure operationalizes the market preconditions that (1) 'the circumstances of time and place' for each economic agent are dispersed and known only to that agent (as in the above voucher market) and (2) agents have a secure property right in the objects of trade and the private gains ('profits') from trade (an airline travel voucher was transferable and redeemable by any bearer). The reader should note that 'profit' is identified as much with the act of buying as with that of selling. This is because 'profit' is the surplus earned by a buyer who buys for less than his willingness-to-pay, just as a seller's 'profit' is the surplus earned when an item is sold for more than the amount for which they are willing to sell. Willingness-to-sell need not have, and usually does not have anything to do with accounting 'cost', or production 'cost', from which one computes accounting profit. Willingness-to-sell, like willingness-to-buy, is determined by the immediate circumstances of each agent. Hence, a passenger might be prepared to pay the regular full fare premium on a first-class ticket for an emergency trip to visit a sick relative. The accountant's concept of profit cannot be applied to the passenger's decision any more than it can be applied to that of a passenger willing to sell a voucher at a deep discount. In what follows I will use the term 'buyer's surplus' or 'seller's surplus' instead of 'profit' to refer to the gains from exchange enjoyed by buyers or sellers because the term 'profit' is so strongly, exclusively and misleadingly associated with selling activities.

Now let us interpret the previously cited fundamental theorem of economics in the context of the experimental design contained in Fig. 1. We note first that the ordered set of seller (buyer) limit prices defines a supply (demand) function (Fig. 1). A supply (demand) function provides a list of the total quantities that sellers (buyers) would be willing to sell (buy) at corresponding hypothetical fixed prices. Neither of these functions is capable of being observed, scientifically, in the field. This is because the postulated limit prices are inherently private and not publicly observable. We could poll every potential seller (buyer) of vouchers in Chicago's O'Hare airport on 20 December 1979 to get each person's reported limit price, but we would have no way of validating the 'observations' thus obtained. Referring to Fig. 1, we see that in my 1956 experiment, sellers (hypothetically) were just willing to sell three units at price 1.25, nine units at 2.75 and so on. Similarly buyers (hypothetically) were just willing to buy four units at 2.50, seven units at 1.75 and so on. If seller 3 is indifferent between selling and not selling at 1.25, and if every seller (buyer) is likewise indifferent at his/her limit price, then any particular unit may not be sold (purchased) at this limit price. One means of dealing with this problem in laboratory markets is to promise to pay a small 'commission', say 5 cents, to each buyer and seller for each unit bought or sold. Thus seller 3 has a small inducement to sell at 1.25 if he can do no better, and buyer 6 has a small inducement to buy at 2.00 if she can do no better.

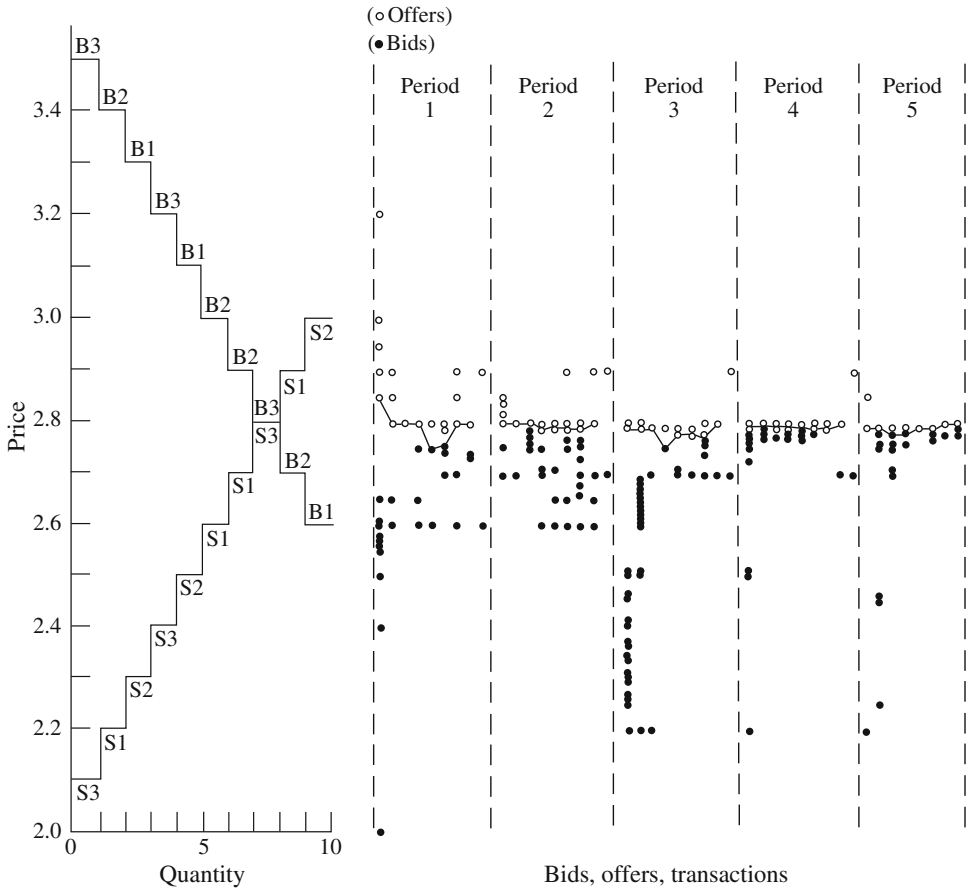
Economic theory defines the competitive equilibrium as the price and corresponding quantity that clears the market; that is, it sets the quantity that sellers are willing to sell equal to the quantity that buyers are willing to buy. This assumes that the subjective cost of transacting is zero; otherwise any units with limit prices equal to the competitive equilibrium price will not exchange.

In Fig. 1 this competitive equilibrium price is 2.00. If the 5 cent 'commission' paid to each trading buyer and seller is sufficient to compensate for any subjective cost of transacting, then buyer 6 and seller 6 will each trade and the competitive equilibrium quantity exchanged will be 6 units. At the competitive equilibrium price, buyer 1 earns a surplus of 3.25–2.00 = 1.25 (plus commission) per period and so on. Total surplus, which measures the maximum possible gains from exchange, or maximum wealth *created* by the existence of the market institution, is 7.50 per period, at the competitive equilibrium.

If by some miracle the competitive equilibrium price and exchange quantity were to prevail in this market, sellers 1–6 would sell, buyers 1–6 would buy, while sellers 7–11 would make no sales and buyers 7–11 would make no purchases. It might be thought that this is unfair – the market should permit some or all of the 'submarginal' buyers (sellers) 7–11 to trade – or that more wealth would be created if there were more than six exchanges. But these interpretations are wrong. By definition, buyer 10 is not willing to pay more than 1.00. Consequently, it is a peculiar notion of fairness to argue that buyer 10 should have as much priority as buyer 1 in obtaining a unit. In the airline voucher example, this would mean that a buyer who is unlikely to redeem a voucher should have the same priority as a buyer who is likely to redeem a voucher. One can imagine a market in which, say, buyer 1 is paired with seller 9 at price 3.00, buyer 2 with seller 8 at price 2.75, and so on with nine units traded. If this were to occur it would mean buyers 7–9, who are less likely to use vouchers, have purchased them, and sellers 7–9, who initially held vouchers, and were more likely to use them than buyers 7–9, have sold their vouchers. Furthermore, this allocation yields additional possible gains from exchange, and is thus *not sustainable*, even if it were thought to be desirable. That is, buyer 9, who bought from seller 1 at price 1.00, could resell the unit to seller 9 (who sold her unit to buyer 2), at price (say) 2.00. Why? Because, by definition a voucher is worth 2.75 to seller 9 and only 1.25 to buyer 9. Similar additional trades can be made by buyers (sellers) 7 and 8. The end result would be that buyers 1–6 and sellers 7–11 would be the terminal holders of vouchers, just as if the competitive equilibrium had been reached initially.

Hence, either the competitive equilibrium prevails, or if inefficient trades occur at dispersed prices, then further 'speculative' gains can be made by some buyers and sellers. If these gains are fully captured the end result is the same allocation as would occur at the competitive equilibrium price and quantity.

Having specified the environment (individual private values) of our experimental market, what remains is to specify an exchange institution. In my 1956 experiment I elected to use trading rules similar to those that characterize trading on the organized stock and commodity exchanges. These markets use the 'double oral auction' procedure. In this institution as soon as the market 'opens' any buyer is free to announce a bid to buy and any seller is free to announce an offer to sell. In the experimental version each bid (offer) is for a single unit. Thus a buyer might say 'buy, 1.00', while a seller might say 'sell, 5.00', and it is understood that the buyer bids 1.00 for a unit and the seller offers to sell one unit for 5.00. Bids and offers are freely announced and can be modified. A contract occurs if any seller accepts the bid of any buyer, or any buyer accepts the offer of any seller. In the simple experimental market, since each participant is a buyer or seller of at most one unit per trading period, the contracting buyer and seller drop out of the market for the remainder of the trading period, but return to the market when a new trading 'day' begins. The experimenter announces the close of each trading period and the opening of the subsequent period, with each trading period timed to extend, say, five minutes. Each contract price is plotted on the right of Fig. 1 for the five trading periods of the experiment. This result was not as expected. The conventional view among economists was that a competitive equilibrium was like a frictionless ideal state which could not be conceived as actually occurring, even approximately. It could be conceived of occurring only in the presence of an abstract 'institution' such as a Walrasian *tâtonnement* or an Edgeworth recontracting procedure. It was for teaching, not believing.



Experimental Methods in Economics, Fig. 2

From Fig. 1 it is evident that in the strict sense the competitive equilibrium was not attained in any period, but the accuracy of the competitive equilibrium theory is easily comparable to that of countless physical processes. Certainly, the data clearly do not support the monopoly, or seller collusion model. The total return to sellers is maximized when four units are sold at price 2.50. Similarly, the monopsony, or buyer collusion model requires four units to exchange at price 1.50.

Since 1956, several hundred experiments using different supply and demand conditions, experienced as well as inexperienced subjects, buyers and sellers with multiple unit trading capacity, a great variation in the numbers of buyers and sellers, and different trading institutions, have established the replicability and robustness of these results. For many years at the University of Arizona and Indiana University we have been using various computerized (the PLATO system) versions of the double 'oral' auction, developed by Arlington Williams, in which participating subjects trade with each other through computer terminals. These experiments establish that the 1956 results are robust with respect to substantial reductions in the number of buyers and sellers. Most such experiments use only four buyers and four sellers, each capable of trading several units. Some have used only two sellers, yet the competitive equilibrium model performs very well under double auction rules. Figure 2 shows the supply and demand design and the market results for a typical experiment in which subjects trade through PLATO computer terminals under computer-monitored double auction rules.

In addition to its antiquarian value, Fig. 1 illustrates the problem of monitoring the rules of a 'manual' experiment. Observe that in period 4 there were seven contracts which are recorded as occurring in the price range between $1.90 and $2.25. This is not possible since there are only six buyers with limit buy prices above $1.90. Either a buyer violated his budget constraint, or the experimenter erred in recording a price in his first experiment. In Fig. 2 there is plotted each contract (an accepted bid if the contract line passes through a 'dot'; an accepted offer if the line passes through a 'circle') and the bids ('dots') and offers ('circles') that preceded each accepted bid or offer. One of the several advantages of computerized experimental markets is that the complete data of the market (all bids, offers, and contracts at their time of execution) are recorded accurately and non-invasively, and all experimental rules are enforced perfectly. In particular the violation of a budget constraint revealed in Fig. 1, which is a perpetual problem with manually executed experiments, is not a problem when trading is perfectly computer monitored.

The rapid convergence shown in Figs. 1 and 2 has not always extended to trading institutions other than the double auction. For example, the 'posted offer' pricing mechanism (associated with most retail markets), in which sellers post take it or leave it non-negotiable prices at the beginning of each period, yields higher prices and less efficient allocations than the double auction. This difference in performance becomes smaller with experienced subjects and with longer trading sequences in a given experiment (Ketcham et al. 1984). Similarly, a comparison of double auction with a sealed bid-offer auction finds the latter to be less efficient and to deviate more from the competitive equilibrium predictions (Smith et al. 1982). Thus, institutions have been demonstrated to make a difference in what we observe. The data and analysis strongly suggest that institutions make a difference because the rules (legal environment) make a difference, and the rules make a difference because they affect individual incentives.

Brief Interpretive History of the Development of Experimental Economics

The two most influential early experimental studies represent the two most primary poles of experimental economics: the study of individual preference (choice) under uncertainty (Mosteller and Nogee 1951) and of market behaviour (Chamberlin 1948). The investigation of uncertainty and preference has focused on the testing of von Neumann–Morgenstern–Savage subjective expected utility theory. Battalio, Kagel and others have pioneered in the testing of the Slutsky–Hicks commodity demand and labour supply preferences using humans (1973) and animals (1975). A series of large-scale field experiments in the 1970s extended the experimental study of individual preference to the measurement of the effect of the negative income tax and other factors on labour supply and to the measurement of the demand for electricity, housing and medical services.

Since the human species has been observed to participate in market exchange for thousands of years, the experimental study of market behaviour is central to economics. Preferences are not directly observable, but preference theory, as an abstract construct, has been *postulated* by economists to be fundamental to the explanation and understanding of market behaviour. In this sense the experimental study of group market behaviour depends upon the study of individual preference behaviour. But this intellectual history should not obscure the fact that the study of markets and the study of preferences need not be construed as inseparable. Adam Smith clearly viewed the human 'propensity to truck, barter and exchange' (and *not* the existence of human preferences) as axiomatic to the scientific study of economic behaviour. Obversely, the work of Battalio and Kagel showing that animals behave as if they had Slutsky–Hicks preferences makes it plain that substitution behaviour is an important cross species characteristic, but that such phenomena need not be associated with market exchange.

A significant feature of Chamberlin's (1948) original work is that it concerned the study of behaviourally complete markets; that is all trades, including purchases as well as sales, were executed by active subject agents. This feature has continued in the subsequent bilateral bargaining experiments of Siegel and Fouraker (1960) and in market experiments (Smith 1962, 1982; Williams and Smith 1984) such as those discussed in section "Markets and Market Experiments". This feature was not present in the early and subsequent experimental oligopoly literature (Hoggatt 1959; Sauermann and Selten 1959; Shubik 1962; Friedman 1963), in which the demand behaviour of buyers was simulated, that is, programmed from a specified demand function conditional on the prices selected in each 'trading' period by the sellers. This simulation of demand behaviour is justified as an intermediate step in testing models of seller price behaviour that assume passive, simple maximizing, demand-revelation behaviour by buyers. But the conclusions of such experimental studies should not be assumed to be applicable, even provisionally, to any observed complete market without first showing that the experimental results are robust with respect to the substitution of subject buyers for simulated buyers.

The Functions of Market Experiments in Microeconomic Analysis

A conceptual framework for clarifying some uses and functions of experiments in microeconomics can be articulated by suitable modification and adaptation (Smith 1982) of the concepts underlying the adjustment process, as in the welfare economics literature (see references to Hurwicz and Reiter in Smith 1982). In this literature a microeconomic environment consists of a list of agents $\{1, \ldots, N\}$, a list of commodities and resources $\{1,\ldots, K\}$and certain characteristics of each agent i, such as the agent's preferences (utility) u_i technological (knowledge) endowment T^i, and commodity endowment w_i .Thus agent i is defined by the triplet of characteristics $E^i = (u^i, T^i, w^i)$ defined on the K-dimensional commodity space. A microeconomic *environment* is defined by the collection $E = (E^1, \ldots, E^N)$ of these characteristics. This collection represents a set of primitive circumstances that condition agents' interaction through institutions. The superscript i, besides identifying a particular agent, also means that these primitive circumstances are in their *nature* private: it is the individual who likes, works, knows and makes.

There can be no such thing as a credible institution-free economics. Institutions define the property right rules by which agents communicate and exchange or transform commodities within the limits and opportunities inherent in the environment, E. Since markets require communication to effect exchange, property rights in messages are as important as property rights in goods and ideas. An institution specifies a language, $M = (M^1, \ldots, M^N)$, consisting of message elements $m = (m^1, \ldots, m^N)$, where M^i is the set of messages that can be sent by agent i (for example, the range of bids that can be sent by a buyer). An institution also defines a set of allocation rules $h = (h^1(m), \ldots, h^N(m))$ and a set of cost imputation rules $c = (c^1(m), \ldots, c^N(m))$ where $h^i(m)$ is the commodity allocation to agent i and $c^i(m)$ is the payment to be made by i, each as a function of the messages sent by all agents. Finally, the institution defines a set of adjustment process rules (assumed to be common to all agents), $g(t_0, t, T)$, consisting of a starting rule, $g(t_0, \cdot, \cdot)$, a transition rule, $g(\cdot, t, \cdot)$, governing the sequencing of messages, and a stopping rule, $g(\cdot, \cdot, T)$, which terminates the exchange of messages and triggers the allocation and cost imputation rules. Each agent's property rights in communication and exchange is thus defined by $I^i = (M^i, h^i(m), c^i(m), g(t_0, t, T))$. A microeconomic *institution* is defined by the collection of these individual property right characteristics, $I = (I^1, \ldots, I^N)$.

A microeconomic *system* is defined by the conjunction of an environment and an institution, $S = (E, I)$. To illustrate a microeconomic system, consider an auction for a single indivisible object such as a painting or an antique vase. Let each of

N agents place an independent, certain, monetary value on the item v_1, . . . , v_N, with agent i knowing his own value, v_i, but having only uncertain (probability distribution) information on the values of others. Thus $E^i = (v^i; p(v), N)$. If the exchange institution is the 'first price' sealed-bid auction, the rules are that all N bidders each submit a single bid any time between the announcement of the auction offering at t_0, and the closing of bids, at T. The item is then awarded to the maker of the highest bid at a price equal to the amount bid. Thus, if the agents are numbered in descending order of the bids, the first price auction institution $I_1 = I_1^1 = \left[h^1(m) = 1, c^1(m) = b_1\right]$ and $I_1^i = \left[h^i(m) = 0, c^i(m) = 0\right], i > 1$, where $m = (b_1, \ldots, b_N)$ consists of all bids tendered. That is, the item is awarded to the high bidder, $i = 1$, who pays b_1, and all others receive and pay nothing. This contrasts with the 'second price' sealed-bid auction $I_2 = \left(I_2^1, \ldots, I_2^N\right)$ in which $I_2^1 = \left[h^1(m) = 1, c^1(m) = b_2\right]$ and $I_2^i = \left[h^i(m) = 0, c^i(m) = 0\right], i > 1$; that is, the highest bidder receives the allocation but pays a price equal to the second highest bid submitted.

Another example is the English or progressive oral auction, whose rules are discussed under the entry auctions (experiments). It should be noted that the 'double oral' auction, used extensively in stock and commodity trading and in the two experimental markets discussed in section "Markets and Market Experiments", is a two-sided generalization of the English auction.

A microeconomic system is activated by the behavioural choices of agents in the set M. In the static, or final outcome, description of an economy, agent *behaviour* can be defined as a function (or correspondence) $m^i = \beta^i(E^i|\ I)$ carrying the characteristics E^i of agent i into a message m^i, conditional upon the property right specifications of the operant institution I. If all exchange-relevant agent characteristics are included in E^i, then $\beta \equiv \beta^i$ for all i. Given the message-sending behaviour of each agent, $\beta(E|\ I)$, the institution determines the outcomes

$$h^i(m) = h^i\left[\beta\left(E^1|\ I\right), \ldots, \beta\left(E^N|\ I\right)\right]$$

and

$$c^i(m) = c^i\left[\beta\left(E^1|\ I\right), \ldots, \beta\left(E^N|\ I\right)\right].$$

Within this framework we see that agents do not choose allocations directly; agents choose messages with institutions determining allocations under the rules that carry messages into allocations. (You cannot choose to 'buy' an auctioned item; you can only choose to raise the standing bid at an English auction or submit a particular bid in a sealed bid auction.) However, the allocation and cost imputation rules may have important incentive effects on behaviour, and therefore messages will in general depend on these rules. Hence, market outcomes will result from the conjunction of institutions' and agents' behaviour.

A proper theory of agents' behaviour allows one to deduce a particular β function based on assumptions about the agent's environment and the institution, and his motivation to act. Auction theory is perhaps the only part of economic theory that is fully institution specific. For example, in the second price sealed bid auction it is a dominant strategy for each agent simply to bid his or her value; that is

$$b^i = \beta\left(E^i|\ I_2\right) = \beta(v_i|\ I_2) = v_i, i = 1, \ldots, N.$$

The resulting outcome is that $b_1 = v_1$ is the winning bid and agent 1 pays the price v_2. Similarly, in the English auction, agent 1 will eventually exclude agent 2 by raising the standing bid to v_2 (or somewhat above), and obtain the item at this price. In the first price auction Vickrey proved that if each agent maximizes expected Surplus $(v_i - b_i)$ in an environment with $P(v) = v$ (the vi are drawn from a constant density on [0, 1]), then we can deduce the noncooperative equilibrium bid function, $b_i = \beta(E^i|\ I_1) = \beta[v_i; P(v), N|\ I_1] = (N - 1)\ v_i|N$ (see the entry on auctions (experiments) for a more complete discussion).

With the above framework it is possible to explicate the roles of theory and experiment, and their relationship, in a progressive research programme (Lakatos 1978) of economic analysis. But to do this we must first ask two questions:

(1) 'Which of the elements of a microeconomic system are not observable?' The nonobservable elements are (i) preferences, (ii) knowledge endowments, and (iii) agent message behaviour, $\beta(E^i|\ I)$. Even if messages are available and recorded, we still cannot observe message behaviour functions because we cannot observe, or vary, preferences. The best we can do with field observations of outcomes is to interpret them in terms of models based on assumptions about preferences (Cobb- Douglas, constant elasticity of substitution, homothetic), knowledge (complete, incomplete, common), and behaviour (cooperative, noncooperative). Any 'tests' of such models must necessarily be joint tests of all of these unobservable elements. More often the econometric exercise is parameter estimation, which is conditional upon these same elements.

(2) 'What would we like to know?' We would like to know enough about how agents' behaviour is affected by alternative environments and institutions so that we can classify them according to the mapping they provide into outcomes. Do some institutions yield Pareto optimal outcomes, and/or stable prices, and, if so, are the results robust with respect to alternative environments?

These two questions together tell us that what we want to know is inaccessible in natural experiments (field data) because key elements of the equation are unobservable and/or cannot be controlled. If laboratory experiments are to help us learn what we want to know, certain precepts that constitute proposed sufficient conditions for a valid controlled microeconomic experiment must be satisfied:

(1) *Non-satiation* (or monotonicity of reward). Subject agents strictly prefer any increase in the reward medium, π; that is $U_i(\pi_i)$ is monotone increasing for all i.
(2) *Saliency.* Agents have the unqualified right to claim rewards that increase (decrease) in the good (bad) outcomes, x_i, in an experiment; the institution of an experiment renders these rewards salient by defining outcomes in terms of the message choices of agents.

In both the field and the laboratory it is the institution that induces value on messages, given each agent's (subjective) value of commodity outcomes. In the laboratory we use a monetary reward function to induce utility value on the abstract accounting outcomes ('commodities') of an experiment. Thus, agent i is given a concave schedule, $v_i(x_i)$, defining the 'redemption value' in dollars for x_i units purchased in an experimental market, and is assured of receiving a net payment equal to $v_i(x_i)$ less the purchase prices of the x_i units in the market. If the x_i units are all purchased at price p (which is the assumption used to derive a hypothetical demand schedule) the agent is paid $\pi_i = V_i(x_i) - px_i$, with utility $u^i(x_i) = U_i(\pi_i(x_i))$. In defining demand it is assumed that the agent directly chooses x_i (that is $x_i = m_i$). Therefore, if i maximizes $u^i(x_i) = U_i[V_i(x_i) - px_i]$, then at a maximum we have $U_i'\left[V_i'(x_i) - p\right] = 0$, giving the demand function $x_i = V_i'^{(-1)}(p)$ if $U_i' > 0$, where $V_i'^{(-1)}$ is the inverse of i's marginal redemption value of x_i units. (The same procedure for a seller using a cost function $C_j(x_j)$ and paying $px_j - C_j(x_j)$ allows one to induce a marginal cost supply of j.) This illustration generalizes easily: if the joint redemption value is $V_i(x_i, y_i)$ for two abstract commodities (x_i, y_i) , $u^i = U_i[V^i(x_i, y_i)]$ induces an indifference map given by the level curves of $V^i(x_i, y_i)$, on (x_i, y_i), with marginal rate of substitution $U_i'V_x^i/U_i'V_y^i = V_x^i/V_y^i$, if $U_i' > 0$. if $V^i(x_i, X)$ the reward function, with x_i a private and X a common (public) outcome good, we are able to control preferences in the study of public good allocation mechanisms, or if

$$X = \sum_{i=1}^{N} x_i$$

we are poised to study allocation with an 'atmospheric' externality (Coursey and Smith 1985).

The first two precepts are sufficient to allow us to assert that we have created a microeconomic system $S = (E, I)$ in the laboratory. But to assure that we have created a controlled microeconomy, we need two additional precepts:

(3) *Dominance*. Own rewards dominate any subjective costs of transacting (or other motivation) in the experimental market.

As with any person, subject agents may have variables other than money in their utility functions. In particular, if there is cognitive and kinesthetic (observe the traders on a Stock Exchange floor) disutility associated with the message-transaction process of the institution, then utility might be better written $U_i(\pi_i, m^i)$. To the extent that this is so we induce a smaller demand on i with the payoff $v_i(x_i)$ than was computed above, and we lose control over preferences. As a practical matter experimentalists think the problem can usually be finessed by using rewards that are large relative to the complexity of the task, and by adopting experimental procedures that reduce complexity (e.g. using the computer to record decisions, perform needed calculations, provide perfect recall, etc.). Another approach, as noted in section "Markets and Market Experiments", is to pay a small commission for each trade to compensate for the subjective transaction costs.

(4) *Privacy*. The subjects in an experiment each receive information only on his/ her own reward schedule.

This precept is used to provide control over interpersonal utilities (payoff externalities). Real people may experience negative or positive utilities from the rewards of others, and to the extent that this occurs we lose control over induced demand, supply and preference functions. Remember that the reward functions have the same role in an experiment that preference functions have in the economy, and the latter preferences are private and non-observable.

If our interest is confined to testing hypotheses from theory, we are done. Precepts (1)–(4) are sufficient to provide rigorous tests of the theorist's ability to model individual and market behaviour. But one naturally asks if replicable results from the laboratory are transferable to field environments. This requires:

(5) *Parallelism*. Propositions about behaviour and/or the performance of institutions that have been tested in one microeconomy (laboratory or field) apply also to other microeconomies (laboratory or field) where similar *ceteris paribus* conditions hold.

Astronomy, meteorology, biology and other sciences use the maintained hypothesis that the same physical laws hold everywhere. Economics postulates that when the environment and institution are the same, behaviour will be the same; that is, behaviour is determined by a relatively austere subset of life's parameters. Whether this is 'true' is an empirical question. Hence, when one experimentalist studies variations on the treatment variables of another it is customary to replicate the earlier work to check parallelism. Similarly, one must design field experiments, or devise econometric models using non-experimental field data, that provide tests of the transferability of experimental results to any particular market in the field. Only in this way can questions of parallelism be answered. They are not answered with speculations about alleged differences between the experimental subject's behaviour and (undefined) 'real world' behaviour. The experimental laboratory *is* a real world, with real people, real institutions, real payoffs and commodities just as real as stock certificates and airline travel vouchers, both of which have utility because of the claim rights they legally bestow on the bearer.

Classifying the Application of Experimental Methods

There are many types of experiments and many fields of economic study to which experimental methods have been applied.

The experimental study of auctions makes the most extensive use of models of individual behaviour based explicitly on the message requirements of the different institutions. This literature provides test comparisons of predicted behaviour, $m^i = \beta^i(E_i|\, I)$, with observations on individual choice, $\hat{m}^i = \beta^i\left(\hat{E}^i \,|\, I\right)$, for given realizations,

$\hat{E}^i$ (such as values, $\hat{v}^i$ where they are assigned at random). The large literature on experimental double auctions makes no such individual comparisons, because the theoretical literature had not yielded tractable models of individual bid- offer behaviour (but recent contributions by Friedman (1984), and Wilson (1984) are providing such models). Here as in most other areas of experimental research the comparisons are between the predicted price–quantity outcomes of static theory (such as competitive, monopoly, and Cournot models), and observed outcomes. But double auctions have been studied (see references in Smith 1982) in a variety of environments; for example, the effect of price floors and ceilings have been examined (see references in Plott 1982). In all cases these studies are making comparisons. In *nomotheoretical* experiments one compares theory and observation, whereas in *nomoempirical* experiments one compares the effect of different institutions and/or environments as a means of documenting replicable empirical 'laws' that may stimulate modelling energy in new directions. The idea that formal theory must precede meaningful observation does not account for most of the historical development of science. *Heuristic* or exploratory experiments that provide empirical probes of new topics and new experimental methods should not be discouraged.

In industrial organization, and antitrust economics, experimental methods have been applied to examine the effects of monopoly, conspiracy, and alleged anticompetitive practices, and to study the concept of natural monopoly and its relation to scale economics, entry cost and the contestable markets hypothesis (see references in Plott 1982; Smith 1982; Coursey et al. 1984).

An important development in the experimental study of allocation processes has been the extension of experimental market methods to majority rule (and other) committee processes, and to market-like group processes for the provision of goods which have public or common outcome characteristics (loosely, public goods). These studies have examined public good allocation under majority (and Roberts') rules for committee including the effect of the agenda (see the references to Fiorina and Plott, and Levine and Plott in Smith 1982), and under compensated unanimity processes suggested by theorists (see the references in Coursey and Smith 1985). Generally, this literature reports substantial experimental support for the theory of majority rule outcomes, the theory of agenda processes (the sequencing of issues for voting decisions), and for incentive compatible models of the provision of public goods.

See Also

- ▶ Allais Paradox
- ▶ Efficient Allocation
- ▶ Preference Reversals

Bibliography

Battalio, R., J. Kagel, R. Winkler, E. Fisher, R. Basmann, and L. Krasner. 1973. A test of consumer demand theory using observations of individual consumer purchases. *Western Economic Journal* 11: 411–428.

Chamberlin, E. 1948. An experimental imperfect market. *Journal of Political Economy* 56: 95–108.

Coursey, D., and V. Smith. 1985. Experimental tests of an allocation mechanism for private, public or externality goods. *Scandinavian Journal of Economics* 86: 468–484.

Coursey, D., M. Isaac, M. Luke, and V. Smith. 1984. Market contestability in the presence of sunk (entry) costs. *RAND Journal of Economics* 15 (1): 69–84.

Friedman, J. 1963. Individual behavior in oligopolistic markets: An experimental study. *Yale Economic Essays* 3: 359–417.

Friedman, D. 1984. On the efficiency of experimental double auction markets. *American Economic Review* 74: 60–72.

Hoggatt, A. 1959. An experimental business game. *Behavioral Science* 4 (3): 192–203.

Kagel, J., R. Battalio, H. Rachlin, L. Green, R. Basmann, and W. Klemm. 1975. Experimental studies of consumer behavior using laboratory animals. *Economic Inquiry* 13 (1): 22–38.

Ketcham, J., V. Smith, and A. Williams. 1984. A comparison of posted-offer and double-auction pricing institutions. *Review of Economic Studies* 51: 595–614.

Lakatos, I. 1978. In *The methodology of scientific research programmes, philosophical papers*, ed. J. Worrall and G. Currie, vol. 1. Cambridge: Cambridge University Press.

Mosteller, F., and P. Nogee. 1951. An experimental measurement of utility. *Journal of Political Economy* 59: 371–404.
Plott, C. 1982. Industrial organization theory and experimental economics. *Journal of Economic Literature* 20: 1485–1527.
Sauermann, H., and R. Selten. 1959. Ein Oligopolexperiment. *Zeitschrift für die Gesamte Staatswissenschaft* 115: 427–471.
Shubik, M. 1962. Some experimental non zero sum games with lack of information about the rules. *Management Science* 81: 215–234.
Siegel, S., and L. Fouraker. 1960. *Bargaining and group decision making*. New York: McGraw-Hill.
Smith, V. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70: 111–137.
Smith, V. 1982. Microeconomic systems as experimental science. *American Economic Review* 72: 923–955.
Smith, V., A. Williams, K. Bratton, and M. Vannoni. 1982. Competitive market institutions: Double auctions versus sealed bid-offer auctions. *American Economic-Review* 72: 58–77.
Williams, A., and Smith, V. 1984. Cyclical double-auction markets with and without speculators. *Journal of Business* 57(1) Pt 1, 1–33.
Wilson, R. (1984, August). *Multilateral exchange* (Working paper No. 7). Stanford: Stanford University.

Experimental Methods in Economics (ii)

Vernon L. Smith

Historically, the method and subject matter of economics have presupposed that it was a non-experimental (or 'field observational') science more like astronomy or meteorology than physics or chemistry. Based on general, introspectively 'plausible', assumptions about human preferences, and about the cost and technology based supply response of producers, economists have sought to understand the functioning of economies, using observations generated by economic outcomes realized over time. The data of the astronomer is of this same type, but it would be wrong to conclude that astronomy and economics are methodologically equivalent. There are two important differences between astronomy and economics which help to illuminate some of the methodological problems of economics. First, based upon parallelism (the maintained hypothesis that the same physical laws hold everywhere), astronomy draws on all the relevant theory from classical mechanics and particle physics – theory which has evolved under rigorous laboratory tests. Traditionally, economists have not had an analogous body of tested behavioural principles that have survived controlled experimental tests, and which can be assumed to apply with insignificant error to the microeconomic behaviour that underpins the observable operations of the economy. Analogously, one might have supposed that there would have arisen an important area of common interest between economics and, say, experimental psychology, similar to that between astronomy and physics, but this has only started to develop in recent years.

Second, the data of astronomy are painstakingly gathered by professional observational astronomers for scientific purposes, and these data are taken seriously (if not always non-controversially) by astrophysicists and cosmologists. Most of the data of economics has been collected by government or private agencies for non-scientific purposes. Hence astronomers are directly responsible for the scientific credibility of their data in a way that economists have not been. In economics, when things appear not to turn out as expected the quality of the data is more likely to be questioned than the relevance and quality of the abstract reasoning. Old theories fade away, not from the weight of falsifying evidence that catalyses theoretical creativity into developing better theory, but from lack of interest, as intellectual energy is attracted to the development of new techniques and to the solution of new puzzles that remain untested.

At approximately the mid–20th century, professional economics began to change with the introduction of the laboratory experiment into economic method. In this embryonic research programme economists (and a psychologist, Sidney Siegel) became directly involved in the design and conduct of experiments to examine propositions implied by economic theories of markets. For the first time this made it possible to introduce

demonstrable knowledge into the economist's attempt to understand markets.

This laboratory approach to economics also brought to the economist direct responsibility for an important source of scientific data generated by controlled processes that can be replicated by other experimentalists. This development invited economic theorists to submit to a new discipline, but also brought an important new discipline and new standards of rigour to the data gathering process itself.

An untested theory is simply a hypothesis. As such it is part of our *self*-knowledge. Science seeks to expand our knowledge of *things* by a process of testing this type of self-knowledge. Much of economic theory can be called, appropriately, 'ecclesiastical theory'; it is accepted (or rejected) on the basis of authority, tradition, or opinion about assumptions, rather than on the basis of having survived a rigorous falsification process that can be replicated.

Interest in the replicability of scientific research stems from a desire to answer the question 'Do you see what I see?'. Replication and control are the two primary means by which we attempt to reduce the error in our common knowledge of economic processes. However, the question 'Do you see what I see?' contains three component questions, recognition of which helps to identify three different senses in which a research study may fail to be replicable:

1. *Do you observe what I observe?* Since economics has traditionally been confined to the analysis of non-experimental data, the answer to this question has been trivially, 'yes'. We observe the same thing because we use the same data. This non-replicability of our traditional data sources has helped to motivate some to turn increasingly to experimental methods. We can say that you have replicated my experiments if you are unable to reject the hypothesis that your experimental data came from the same population as mine. This means that the experimenter, his/her subjects, and/or procedures are not significant treatment variables.
2. *Do you interpret what we observe as I interpret it?* Given that we both observe the same, or replicable data, do we put the same interpretation on these data? The interpretation of observations requires theory (either formal or informal), or at least an empirical interpretation of the theory in the context that generated the data. Theory usually requires empirical interpretation either because (i) the theory is not developed directly in terms of what can be observed (e.g. the theory may assume risk aversion which is not directly observable), or (ii) the data were not collected for the purpose of testing, or estimating the parameters of a theory. Consequently, failure to replicate may be due to differences in interpretation which result from different meanings being ascribed to the theory. Thus two researchers may apply different transformations to raw field data (e.g. different adjustments for the effect of taxes), so that the results are not replicable because their theory interpretations differ.
3. *Do you conclude what I conclude from our interpretation?* The conclusions reached in two different research studies may be different even though the data and their interpretation are the same. In economics this is most often due to different model specifications. This problem is inherent in non-experimental methodologies in which, at best, one usually can estimate only the parameters of a prespecified model and cannot credibly test one model or theory against another. An example is the question of whether the Phillips' curve constitutes a behavioural trade-off between the rates of inflation and unemployment, or represents an equilibrium association without causal significance.

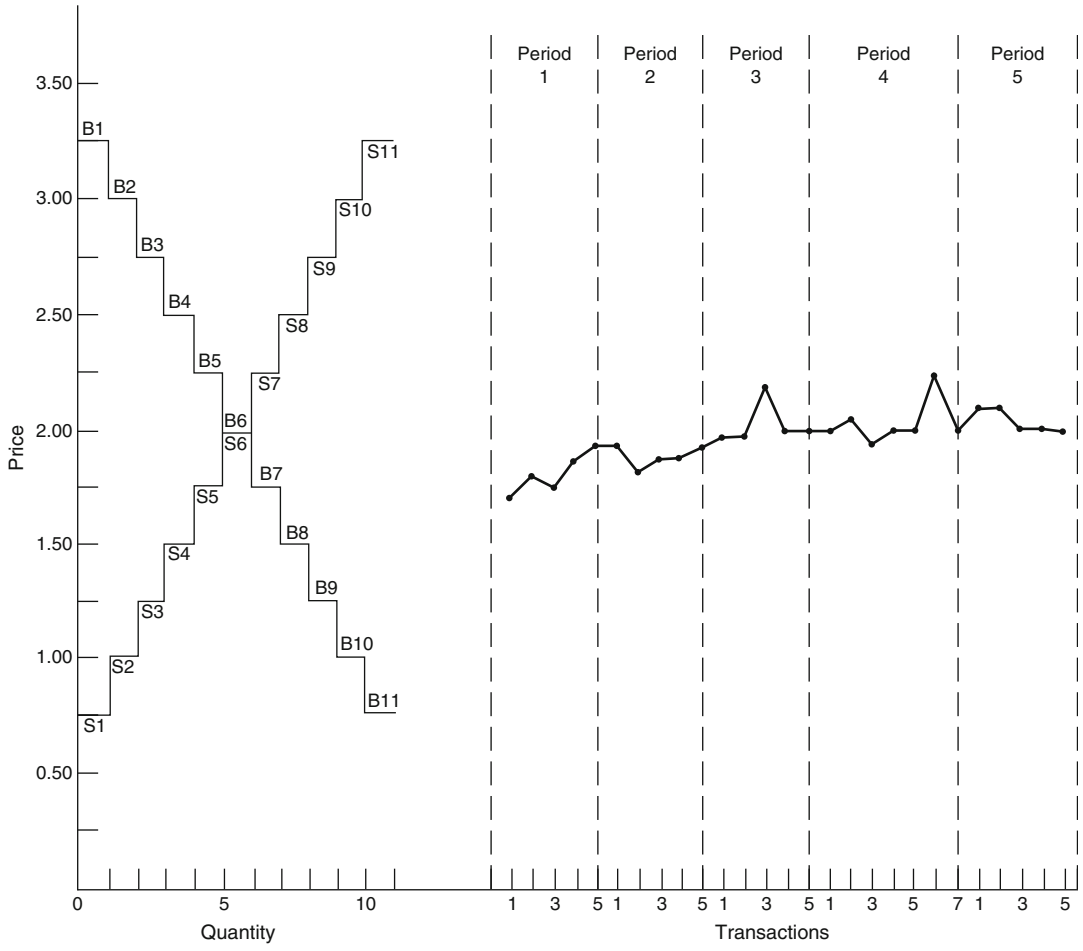
Markets and Market Experiments

Markets and how they function constitute the core of any economic system, whether it is highly decentralized – popularly, a 'capitalistic' system, or highly centralized – popularly, a 'planned' system. This is true for the decentralized economy because markets are the spontaneous institutions of exchange that use prices to guide resource allocation and human economic action. It is true for the centralized economy because in such

economies markets always exist or arise in legal form (private agriculture in Russia) and clandestine or illegal form (barter, bribery, the trading of favours, and underground exchange in Russia, Poland and elsewhere). Markets arise spontaneously in all cultures in response to the human desire for betterment (to 'profit') through exchange. Where the commodity or service is illegal (prostitution, gambling, the sale of liquor under Prohibition or of marijuana, cocaine, etc.) the result is not to prevent exchange, but to raise the risk and therefore the costs of exchange. This is because enforcement is itself costly, and it is never economical for the authorities (whether Soviet or American) even to approximate perfect enforcement. The spontaneity with which markets arise is perhaps no better illustrated than when (1979–1980) US airlines for promotional purposes issued travel vouchers to their passengers. One of these vouchers could be redeemed by the bearer as a cash substitute in the purchase of new airline tickets. Consequently vouchers were of value to future passengers. Furthermore, since (as Hayek would say) the 'circumstances of time and place' for the potential redemption of vouchers were different for different individuals, there existed the preconditions for the active voucher market that was soon observed in all busy airports. Current passengers with vouchers who were unlikely to be travelling again soon held an asset worth less to themselves than to others who were more certain of their future or impending travel plans. The resulting market established prices that were discounts from the redemption or 'face' value of vouchers. Sellers who were unlikely to be able to redeem their vouchers preferred to sell them at a discount for cash. Buyers who were reasonably sure of their travel plans could save money by purchasing vouchers at a discount. Thus the welfare of every active buyer and seller increased via this market. Without a market, many – perhaps most – vouchers would not have been exercised and would thus have been 'wasted'.

The previous paragraph illustrates a fundamental hypothesis (theorem) of economics: the ('competitive') market process yields welfare improving (and, under certain limiting ideal conditions, welfare maximizing) outcomes. But is the hypothesis 'true', or at least very probably true? (Lakatos (1978) would correctly ask 'Has it led to an empirically progressive research programme?') I think it is 'true', but how do I know this? Do you see what I see? A Marxist does not see what I see in the above interpretation of a market. The young student studying economics does not see what I see, although if they continue to study economics eventually they (predictably) come to see what I see (or, at least, they say they do). Is this because we have inadvertently brainwashed them? The gasoline consumer does not see what I see. They see themselves in a *zero* sum game with an oil company: any increase in price merely redistributes wealth from the consumer to the company, which is not 'fair' since the company is richer. What I see in a market is a *positive* sum game yielding gains from exchange, which constitutes the fundamental mechanism for creating, not merely redistributing wealth. The traditional method by which the economist gets others to see this 'true' function of markets is by logical arguments (suppose it were not true, then . . .), examples, and 'observations', such as are contained in my description of the voucher market, in which what is 'observed' is hortatively described and interpreted in terms of the hypothesis itself. But if this knowledge of the function of markets is 'true', can it be demonstrated? Experimentalists claim that laboratory experiments can provide a uniquely important technique of demonstration for supplementing the theoretical interpretation of field observations.

I conducted my first experiment in the spring of 1956. Since then hundreds of similar, as well as environmentally richer experiments have been conducted by myself and by others. In 1956, my introductory economics class consisted of 22 science and engineering students, and although this might not have been the 'large number' traditionally thought to have been necessary to yield a competitive market, I though it was large enough for a practice run to initiate a research programme capable of falsifying the standard theory. I conducted the experiment before lecturing on the theory and 'behaviour' of markets in class so as not to 'contaminate' the sample. The 22 subjects



Experimental Methods in Economics (ii), Fig. 1

were each assigned one card from a well-shuffled deck of 11 white and 11 yellow cards. The white cards identified the sellers, and the yellow cards identified the buyers. Each white card carried a price, known only to that seller, which represented that seller's minimum selling price for one unit, and each yellow card identified a price, known only to that buyer, representing that buyer's maximum buying price for one unit. On the left of Fig. 1 is listed these so-called 'limit' prices, identified by buyer, B1, B2 etc. (in descending order, D) and by seller, S1, S2 etc. (in ascending order, S). To keep things simple and well controlled each buyer (seller) was informed that he/she was a buyer (seller) of at most one unit of the item in each of several trading periods. Thus demand, D(supply, S) was 'renewed' in each trading period as a steady state flow, with no carry-over in unsatisfied demand (or unsold stock), from one period to the next. In the airline voucher example, imagine the vouchers being issued, followed by trading; the vouchers then expire, new vouchers are issued, traded and so on. In the experiment, suppose real motivation is provided by promising to pay (in cash) to each buyer the difference between that buyer's assigned limit buying price and the price actually paid in each period that a unit is purchased in the market. Thus suppose seller 5 sells their unit to buyer 2 at the price 2.25. Then buyer 2 earns a 'profit' of $0.75 from this exchange. In this way we induce on each buyer a value (or hypothesized

willingness-to-pay) equal to the assigned limit buy price. Similarly, suppose each seller is paid the difference between that seller's actual sales price and assigned limit price ('cost', or willingness-to-sell) in each trading period that a unit is sold. Thus in the previous exchange example, seller 5 earns \$0.50 from the transaction.

This experimental procedure operationalizes the market preconditions that (1) 'the circumstances of time and place' for each economic agent are dispersed and known only to that agent (as in the above voucher market) and (2) agents have a secure property right in the objects of trade and the private gains ('profits') from trade (an airline travel voucher was transferable and redeemable by any bearer). The reader should note that 'profit' is identified as much with the act of buying as with that of selling. This is because 'profit' is the surplus earned by a buyer who buys for less than his willingness-to-pay, just as a seller's 'profit' is the surplus earned when an item is sold for more than the amount for which they are willing to sell. Willingness-to-sell need not have, and usually does not have anything to do with accounting 'cost', or production 'cost', from which one computes accounting profit. Willingness-to-sell, like willingness-to-buy, is determined by the immediate circumstances of each agent. Hence, a passenger might be prepared to pay the regular full fare premium on a first-class ticket for an emergency trip to visit a sick relative. The accountant's concept of profit cannot be applied to the passenger's decision any more than it can be applied to that of a passenger willing to sell a voucher at a deep discount. In what follows I will use the term 'buyer's surplus' or 'seller's surplus' instead of 'profit' to refer to the gains from exchange enjoyed by buyers or sellers because the term 'profit' is so strongly, exclusively and misleadingly associated with selling activities.

Now let us interpret the previously cited fundamental theorem of economics in the context of the experimental design contained in Fig. 1. We note first that the ordered set of seller (buyer) limit prices defines a supply (demand) function (Fig. 1). A supply (demand) function provides a list of the total quantities that sellers (buyers) would be willing to sell (buy) at corresponding hypothetical fixed prices. Neither of these functions is capable of being observed, scientifically, in the field. This is because the postulated limit prices are inherently private and not publicly observable. We could poll every potential seller (buyer) of vouchers in Chicago's O'Hare airport on 20 December 1979 to get each person's reported limit price, but we would have no way of validating the 'observations' thus obtained. Referring to Fig. 1, we see that in my 1956 experiment, sellers (hypothetically) were just willing to sell three units at price 1.25, nine units at 2.75 and so on. Similarly buyers (hypothetically) were just willing to buy four units at 2.50, seven units at 1.75 and so on. If seller 3 is indifferent between selling and not selling at 1.25, and if every seller (buyer) is likewise indifferent at his/her limit price, then any particular unit may not be sold (purchased) at this limit price. One means of dealing with this problem in laboratory markets is to promise to pay a small 'commission', say 5 cents, to each buyer and seller for each unit bought or sold. Thus seller 3 has a small inducement to sell at 1.25 if he can do no better, and buyer 6 has a small inducement to buy at 2.00 if she can do no better.

Economic theory defines the competitive equilibrium as the price and corresponding quantity that clears the market; that is, it sets the quantity that sellers are willing to sell equal to the quantity that buyers are willing to buy. This assumes that the subjective cost of transacting is zero; otherwise any units with limit prices equal to the competitive equilibrium price will not exchange. In Fig. 1 this competitive equilibrium price is 2.00. If the 5 cent 'commission' paid to each trading buyer and seller is sufficient to compensate for any subjective cost of transacting, then buyer 6 and seller 6 will each trade and the competitive equilibrium quantity exchanged will be 6 units. At the competitive equilibrium price, buyer 1 earns a surplus of $3.25 - 2.00 = 1.25$ (plus commission) per period and so on. Total surplus, which measures the maximum possible gains from exchange, or maximum wealth *created* by the existence of the market institution, is 7.50 per period, at the competitive equilibrium.

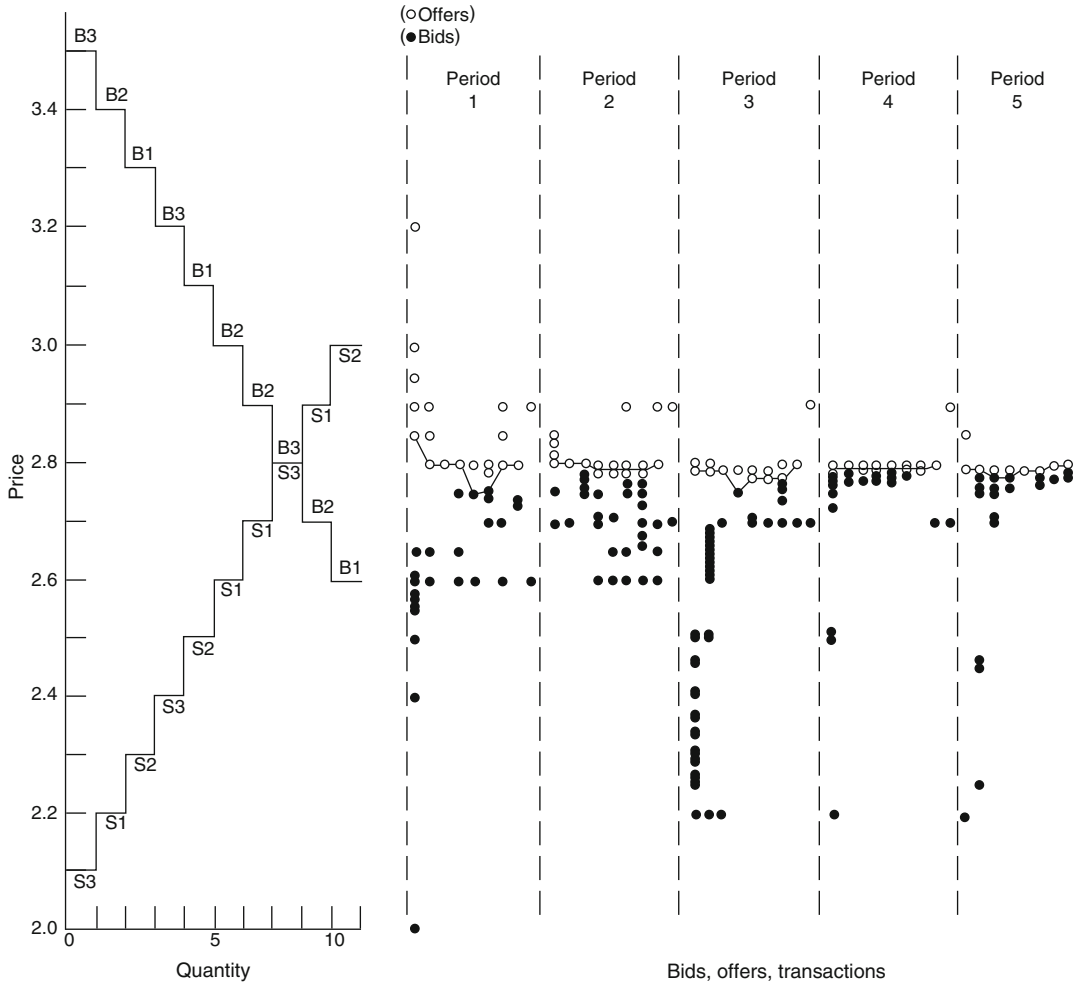
If by some miracle the competitive equilibrium price and exchange quantity were to prevail in this market, sellers 1–6 would sell, buyers 1–6 would buy, while sellers 7–11 would make no sales and buyers 7–11 would make no purchases. It might be thought that this is unfair – the market should permit some or all of the 'submarginal' buyers (sellers) 7–11 to trade – or that more wealth would be created if there were more than six exchanges. But these interpretations are wrong. By definition, buyer 10 is not willing to pay more than 1.00. Consequently, it is a peculiar notion of fairness to argue that buyer 10 should have as much priority as buyer 1 in obtaining a unit. In the airline voucher example, this would mean that a buyer who is unlikely to redeem a voucher should have the same priority as a buyer who is likely to redeem a voucher. One can imagine a market in which, say, buyer 1 is paired with seller 9 at price 3.00, buyer 2 with seller 8 at price 2.75, and so on with nine units traded. If this were to occur it would mean buyers 7–9, who are less likely to use vouchers, have purchased them, and sellers 7–9, who initially held vouchers, and were more likely to use them than buyers 7–9, have sold their vouchers. Furthermore, this allocation yields additional possible gains from exchange, and is thus *not sustainable*, even if it were thought to be desirable. That is, buyer 9, who bought from seller 1 at price 1.00, could resell the unit to seller 9 (who sold her unit to buyer 2), at price (say) 2.00. Why? Because, by definition a voucher is worth 2.75 to seller 9 and only 1.25 to buyer 9. Similar additional trades can be made by buyers (sellers) 7 and 8. The end result would be that buyers 1–6 and sellers 7–11 would be the terminal holders of vouchers, just as if the competitive equilibrium had been reached initially.

Hence, either the competitive equilibrium prevails, or if inefficient trades occur at dispersed prices, then further 'speculative' gains can be made by some buyers and sellers. If these gains are fully captured the end result is the same allocation as would occur at the competitive equilibrium price and quantity.

Having specified the environment (individual private values) of our experimental market, what remains is to specify an exchange institution. In my 1956 experiment I elected to use trading rules similar to those that characterize trading on the organized stock and commodity exchanges. These markets use the 'double oral auction' procedure. In this institution as soon as the market 'opens' any buyer is free to announce a bid to buy and any seller is free to announce an offer to sell. In the experimental version each bid (offer) is for a single unit. Thus a buyer might say 'buy, 1.00', while a seller might say 'sell, 5.00', and it is understood that the buyer bids 1.00 for a unit and the seller offers to sell one unit for 5.00. Bids and offers are freely announced and can be modified. A contract occurs if any seller accepts the bid of any buyer, or any buyer accepts the offer of any seller. In the simple experimental market, since each participant is a buyer or seller of at most one unit per trading period, the contracting buyer and seller drop out of the market for the remainder of the trading period, but return to the market when a new trading 'day' begins. The experimenter announces the close of each trading period and the opening of the subsequent period, with each trading period timed to extend, say, five minutes. Each contract price is plotted on the right of Fig. 1 for the five trading periods of the experiment. This result was not as expected. The conventional view among economists was that a competitive equilibrium was like a frictionless ideal state which could not be conceived as actually occurring, even approximately. It could be conceived of occurring only in the presence of an abstract 'institution' such as a Walrasian *tâtonnement* or an Edgeworth recontracting procedure. It was for teaching, not believing.

From Fig. 1 it is evident that in the strict sense the competitive equilibrium was not attained in any period, but the accuracy of the competitive equilibrium theory is easily comparable to that of countless physical processes. Certainly, the data clearly do not support the monopoly, or seller collusion model. The total return to sellers is maximized when four units are sold at price 2.50. Similarly, the monopsony, or buyer collusion model requires four units to exchange at price 1.50.

Since 1956, several hundred experiments using different supply and demand conditions,



Experimental Methods in Economics (ii), Fig. 2

experienced as well as inexperienced subjects, buyers and sellers with multiple unit trading capacity, a great variation in the numbers of buyers and sellers, and different trading institutions, have established the replicability and robustness of these results. For many years at the University of Arizona and Indiana University we have been using various computerized (the PLATO system) versions of the double 'oral' auction, developed by Arlington Williams, in which participating subjects trade with each other through computer terminals. These experiments establish that the 1956 results are robust with respect to substantial reductions in the number of buyers and sellers. Most such experiments use only four buyers and four sellers, each capable of trading several units. Some have used only two sellers, yet the competitive equilibrium model performs very well under double auction rules. Figure 2 shows the supply and demand design and the market results for a typical experiment in which subjects trade through PLATO computer terminals under computer-monitored double auction rules.

In addition to its antiquarian value, Fig. 1 illustrates the problem of monitoring the rules of a 'manual' experiment. Observe that in period 4 there were seven contracts which are recorded as occurring in the price range between $1.90 and $2.25. This is not possible since there are only six

buyers with limit buy prices above $1.90. Either a buyer violated his budget constraint, or the experimenter erred in recording a price in his first experiment. In Fig. 2 there is plotted each contract (an accepted bid if the contract line passes through a 'dot'; an accepted offer if the line passes through a 'circle') and the bids ('dots') and offers ('circles') that preceded each accepted bid or offer. One of the several advantages of computerized experimental markets is that the complete data of the market (all bids, offers, and contracts at their time of execution) are recorded accurately and non-invasively, and all experimental rules are enforced perfectly. In particular the violation of a budget constraint revealed in Fig. 1, which is a perpetual problem with manually executed experiments, is not a problem when trading is perfectly computer monitored.

The rapid convergence shown in Figs. 1 and 2 has not always extended to trading institutions other than the double auction. For example, the 'posted offer' pricing mechanism (associated with most retail markets), in which sellers post take it or leave it non-negotiable prices at the beginning of each period, yields higher prices and less efficient allocations than the double auction. This difference in performance becomes smaller with experienced subjects and with longer trading sequences in a given experiment (Ketcham et al. 1984). Similarly, a comparison of double auction with a sealed bid-offer auction finds the latter to be less efficient and to deviate more from the competitive equilibrium predictions (Smith et al. 1982). Thus, institutions have been demonstrated to make a difference in what we observe. The data and analysis strongly suggest that institutions make a difference because the rules (legal environment) make a difference, and the rules make a difference because they affect individual incentives.

Brief Interpretive History of the Development of Experimental Economics

The two most influential early experimental studies represent the two most primary poles of experimental economics: the study of individual preference (choice) under uncertainty (Mosteller and Nogee 1951) and of market behaviour (Chamberlin 1948). The investigation of uncertainty and preference has focused on the testing of von Neumann–Morgenstern–Savage subjective expected utility theory. Battalio, Kagel and others have pioneered in the testing of the Slutsky–Hicks commodity demand and labour supply preferences using humans (1973) and animals (1975). A series of large-scale field experiments in the 1970s extended the experimental study of individual preference to the measurement of the effect of the negative income tax and other factors on labour supply and to the measurement of the demand for electricity, housing and medical services.

Since the human species has been observed to participate in market exchange for thousands of years, the experimental study of market behaviour is central to economics. Preferences are not directly observable, but preference theory, as an abstract construct, has been *postulated* by economists to be fundamental to the explanation and understanding of market behaviour. In this sense the experimental study of group market behaviour depends upon the study of individual preference behaviour. But this intellectual history should not obscure the fact that the study of markets and the study of preferences need not be construed as inseparable. Adam Smith clearly viewed the human 'propensity to truck, barter and exchange' (and *not* the existence of human preferences) as axiomatic to the scientific study of economic behaviour. Obversely, the work of Battalio and Kagel showing that animals behave as if they had Slutsky–Hicks preferences makes it plain that substitution behaviour is an important cross species characteristic, but that such phenomena need not be associated with market exchange.

A significant feature of Chamberlin's (1948) original work is that it concerned the study of behaviourally complete markets; that is all trades, including purchases as well as sales, were executed by active subject agents. This feature has continued in the subsequent bilateral bargaining experiments of Siegel and Fouraker (1960) and in market experiments (Smith 1962, 1982; Williams and Smith 1984) such as those discussed in

section, "Markets and Market Experiments". This feature was not present in the early and subsequent experimental oligopoly literature (Hoggatt 1959; Sauermann and Selten 1959; Shubik 1962; Friedman 1963), in which the demand behaviour of buyers was simulated, that is, programmed from a specified demand function conditional on the prices selected in each 'trading' period by the sellers. This simulation of demand behaviour is justified as an intermediate step in testing models of seller price behaviour that assume passive, simple maximizing, demand-revelation behaviour by buyers. But the conclusions of such experimental studies should not be assumed to be applicable, even provisionally, to any observed complete market without first showing that the experimental results are robust with respect to the substitution of subject buyers for simulated buyers.

The Functions of Market Experiments in Microeconomic Analysis

A conceptual framework for clarifying some uses and functions of experiments in microeconomics can be articulated by suitable modification and adaptation (Smith 1982) of the concepts underlying the adjustment process, as in the welfare economics literature (see references to Hurwicz and Reiter in Smith 1982). In this literature a microeconomic environment consists of a list of agents $\{1, \ldots, N\}$, a list of commodities and resources $\{1, \ldots, K\}$, and certain characteristics of each agent i, such as the agent's preferences (utility) u_i, technological (knowledge) endowment T^i, and commodity endowment w_i. Thus agent i is defined by the triplet of characteristics $E^i = (u^i, T^i, w^i)$ defined on the K-dimensional commodity space. A microeconomic *environment* is defined by the collection $E = (E^1, \ldots, E^N)$ of these characteristics. This collection represents a set of primitive circumstances that condition agents' interaction through institutions. The superscript i, besides identifying a particular agent, also means that these primitive circumstances are in their *nature* private: it is the individual who likes, works, knows and makes.

There can be no such thing as a credible institution-free economics. Institutions define the property right rules by which agents communicate and exchange or transform commodities within the limits and opportunities inherent in the environment, E. Since markets require communication to effect exchange, property rights in messages are as important as property rights in goods and ideas. An institution specifies a language, $M = (M^1, \ldots, M^N)$, consisting of message elements $m = (m^1, \ldots, m^N)$, where M^i is the set of messages that can be sent by agent i (for example, the range of bids that can be sent by a buyer). An institution also defines a set of allocation rules $h = (h^1(m),\ldots, h^N(m))$ and a set of cost imputation rules $c = (c^1(m), \ldots, c^N(m))$, where h^i(m) is the commodity allocation to agent i and $c^i(m)$ is the payment to be made by i, each as a function of the messages sent by all agents. Finally, the institution defines a set of adjustment process rules (assumed to be common to all agents), $g(t_0, t, T)$, consisting of a starting rule, $g(t_0, \cdot, \cdot)$, a transition rule, $g(\cdot, t, \cdot)$, governing the sequencing of messages, and a stopping rule, $g(\cdot, \cdot, T)$, which terminates the exchange of messages and triggers the allocation and cost imputation rules. Each agent's property rights in communication and exchange is thus defined by $I^i = (M^i, h^i(m), c^i(m), g(t_0, t, T))$. A microeconomic *institution* is defined by the collection of these individual property right characteristics, $I = (I^1,\ldots,I^N)$.

A microeconomic *system* is defined by the conjunction of an environment and an institution, $S = (E, I)$. To illustrate a microeconomic system, consider an auction for a single indivisible object such as a painting or an antique vase. Let each of N agents place an independent, certain, monetary value on the item $v_1, \ldots, v_N$, with agent i knowing his own value, v_i, but having only uncertain (probability distribution) information on the values of others. Thus $E^i = (v^i; P(v), N)$. If the exchange institution is the 'first price' sealed-bid auction, the rules are that all N bidders each submit a single bid any time between the announcement of the auction offering at t_0, and the closing of bids, at T. The item is then awarded to the maker of the highest bid at a price equal to the amount bid. Thus, if the agents are numbered in

descending order of the bids, the first price auction institution $I_1 = (I_1^1 = [h^1(m) = 1, c^1(m) = b_1]$ and $I_1^i = [h^i(m) = 0, c^i(m) = 0], i > 1$, where $m = (b_1, \ldots, b_N)$ consists of all bids tendered. That is, the item is awarded to the high bidder, $i = 1$, who pays b_1, and all others receive and pay nothing. This contrasts with the 'second price' sealed-bid auction $I_2 = (I_2^1, \ldots, I_2^N)$ in which $I_2^1 = [h^i(m) = 1, c^i(m) = b_2]$ and $I_2^2 = [h^i(m) = 0, c^i(m) = 0], i > 1$; that is, the highest bidder receives the allocation but pays a price equal to the second highest bid submitted.

Another example is the English or progressive oral auction, whose rules are discussed under the entry AUCTIONS. It should be noted that the 'double oral' auction, used extensively in stock and commodity trading and in the two experimental markets discussed in section, "Markets and Market Experiments", is a two-sided generalization of the English auction.

A microeconomic system is activated by the behavioural choices of agents in the set M. In the static, or final outcome, description of an economy, agent *behaviour* can be defined as a function (or correspondence) $m^i = \beta^i(E^i|I)$ carrying the characteristics E^i of agent i into a message m^i, conditional upon the property right specifications of the operant institution I. If all exchange-relevant agent characteristics are included in E^i, then $\beta \equiv \beta^i$ for all i. Given the message-sending behaviour of each agent, $\beta(E \mid I)$, the institution determines the outcomes

$$h^i(m) = h^i[\beta E^1 I), \ldots, \beta(E^N I)]$$

and

$$c^i(m) = c^i[\beta E^1|I), \ldots, \beta(E^N|I)].$$

Within this framework we see that agents do not choose allocations directly; agents choose messages with institutions determining allocations under the rules that carry messages into allocations. (You cannot choose to 'buy' an auctioned item; you can only choose to raise the standing bid at an English auction or submit a particular bid in a sealed bid auction.) However, the allocation and cost imputation rules may have important incentive effects on behaviour, and therefore messages will in general depend on these rules. Hence, market outcomes will result from the conjunction of institutions' and agents' behaviour.

A proper theory of agents' behaviour allows one to deduce a particular β function based on assumptions about the agent's environment and the institution, and his motivation to act. Auction theory is perhaps the only part of economic theory that is fully institution specific. For example, in the second price sealed bid auction it is a dominant strategy for each agent simply to bid his or her value; that is

$$b^i = \beta[E^i|I_2) = \beta(v_i|I_2) = v_i, \quad i = 1, \ldots, N.$$

The resulting outcome is that $b_1 = v_1$ is the winning bid and agent 1 pays the price v_2. Similarly, in the English auction, agent 1 will eventually exclude agent 2 by raising the standing bid to v_2 (or somewhat above), and obtain the item at this price. In the first price auction Vickrey proved that if each agent maximizes expected surplus $(v_i - b_i)$ in an environment with $P(v) = v$ (the v_i are drawn from a constant density on [0, 1]), then we can deduce the noncooperative equilibrium bid function, $b_i = \beta(E^i \mid I_1) = \beta[v_i; P(v), N| I_1] = (N - 1)\, v_i|N$ (see the entry on AUCTIONS for a more complete discussion).

With the above framework it is possible to explicate the roles of theory and experiment, and their relationship, in a progressive research programme (Lakatos 1978) of economic analysis. But to do this we must first ask two questions:

1. 'Which of the elements of a microeconomic system are not observable?' The non-observable elements are (i) preferences, (ii) knowledge endowments, and (iii) agent message behaviour, $\beta(E^i|I)$. Even if messages are available and recorded, we still cannot observe message behaviour functions because we cannot observe, or vary, preferences. The best we can do with field observations of outcomes is to interpret them in terms of models

based on assumptions about preferences (Cobb-Douglas, constant elasticity of substitution, homothetic), knowledge (complete, incomplete, common), and behaviour (cooperative, noncooperative). Any 'tests' of such models must necessarily be joint tests of all of these unobservable elements. More often the econometric exercise is parameter estimation, which is conditional upon these same elements.

2. 'What would we like to know?' We would like to know enough about how agents' behaviour is affected by alternative environments and institutions so that we can classify them according to the mapping they provide into outcomes. Do some institutions yield Pareto optimal outcomes, and/or stable prices, and, if so, are the results robust with respect to alternative environments?

These two questions together tell us that what we want to know is inaccessible in natural experiments (field data) because key elements of the equation are unobservable and/or cannot be controlled. If laboratory experiments are to help us learn what we want to know, certain precepts that constitute proposed sufficient conditions for a valid controlled microeconomic experiment must be satisfied:

1. *Non-satiation* (or monotonicity of reward). Subject agents strictly prefer any increase in the reward medium, π ; that is U_i (π_i) is monotone increasing for all i.
2. *Saliency.* Agents have the unqualified right to claim rewards that increase (decrease) in the good (bad) outcomes, x_i, in an experiment; the institution of an experiment renders these rewards salient by defining outcomes in terms of the message choices of agents.

In both the field and the laboratory it is the institution that induces value on messages, given each agent's (subjective) value of commodity outcomes. In the laboratory we use a monetary reward function to induce utility value on the abstract accounting outcomes ('commodities') of an experiment. Thus, agent i is given a concave schedule, $V_i(x_i)$, defining the 'redemption value' in dollars for x_i units purchased in an experimental market, and is assured of receiving a net payment equal to $V_i(x_i)$ less the purchase prices of the x_i units in the market. If the x_i units are all purchased at price p (which is the assumption used to derive a hypothetical demand schedule) the agent is paid $\pi_i = V_i(x_i) - px_i$, with utility $u^i(x) = U(\pi(x_i))$. In defining demand it is assumed that the agent directly chooses x_i (that is $x_i = m_i$). Therefore, if i maximizes $u^i(x_i) = U_i[V_i(x_i) - px]$, then at a maximum we have $U_i' \cdot [V_i'(x_i) - p] = 0$, giving the demand function $x_i = V_i^{'(-1)}(p)$ *if* $U_i' > 0$, where $V_i^{'(-1)}$ is the inverse of i's marginal redemption value of x_i units. (The same procedure for a seller using a cost function $C_j(x_j)$ and paying $px_j - C_j(x_j)$ allows one to induce a marginal cost supply of j.) This illustration generalizes easily: if the joint redemption value is $V_i(x_i, y_i)$ for two abstract commodities (x_i, y_i), $u^i = U_i[V^i(x_i, y_i)]$ induces an indifference map given by the level curves of $V^i(x_i, y_i)$, on (x_i, y_i), with marginal rate of substitution $U_i'V_x^i/U_i'V_y^i = V_x^i/V_y^i$, if $U_i' > 0$. If $Vi(x_i, X)$ the reward function, with x_i a private and X a common (public) outcome good, we are able to control preferences in the study of public good allocation mechanisms, or if

$$X = \sum_{i=1}^{N} x_i$$

we are poised to study allocation with an 'atmospheric' externality (Coursey and Smith 1985).

The first two precepts are sufficient to allow us to assert that we have created a microeconomic sysem $S = (E, I)$ in the laboratory. But to assure that we have created a controlled microeconomy, we need two additional precepts:

3. *Dominance*. Own rewards dominate any subjective costs of transacting (or other motivation) in the experimental market.

As with any person, subject agents may have variables other than money in their utility functions. In particular, if there is cognitive and kinesthetic (observe the traders on a Stock Exchange

floor) disutility associated with the message-transaction process of the institution, then utility might be better written $U_i(\pi_i, m_i)$. To the extent that this is so we induce a smaller demand on i with the payoff $V_i(x_i)$ than was computed above, and we lose control over preferences. As a practical matter experimentalists think the problem can usually be finessed by using rewards that are large relative to the complexity of the task, and by adopting experimental procedures that reduce complexity (e.g. using the computer to record decisions, perform needed calculations, provide perfect recall, etc.). Another approach, as noted in section, "Markets and Market Experiments", is to pay a small commission for each trade to compensate for the subjective transaction costs.

4. *Privacy.* The subjects in an experiment each receive information only on his/her own reward schedule.

This precept is used to provide control over interpersonal utilities (payoff externalities). Real people may experience negative or positive utilities from the rewards of others, and to the extent that this occurs we lose control over induced demand, supply and preference functions. Remember that the reward functions have the same role in an experiment that preference functions have in the economy, and the latter preferences are private and non-observable.

If our interest is confined to testing hypotheses from theory, we are done. Precepts (1)–(4) are sufficient to provide rigorous tests of the theorist's ability to model individual and market behaviour. But one naturally asks if replicable results from the laboratory are transferable to field environments. This requires.

5. *Parallelism*. Propositions about behaviour and/or the performance of institutions that have been tested in one microeconomy (laboratory or field) apply also to other microeconomies (laboratory or field) where similar *ceteris paribus* conditions hold.

Astronomy, meteorology, biology and other sciences use the maintained hypothesis that the same physical laws hold everywhere. Economics postulates that when the environment and institution are the same, behaviour will be the same; that is, behaviour is determined by a relatively austere subset of life's parameters. Whether this is 'true' is an empirical question. Hence, when one experimentalist studies variations on the treatment variables of another it is customary to replicate the earlier work to check parallelism. Similarly, one must design field experiments, or devise econometric models using non-experimental field data, that provide tests of the transferability of experimental results to any particular market in the field. Only in this way can questions of parallelism be answered. They are not answered with speculations about alleged differences between the experimental subject's behaviour and (undefined) 'real world' behaviour. The experimental laboratory *is* a real world, with real people, real institutions, real payoffs and commodities just as real as stock certificates and airline travel vouchers, both of which have utility because of the claim rights they legally bestow on the bearer.

Classifying the Application of Experimental Methods

There are many types of experiments and many fields of economic study to which experimental methods have been applied.

The experimental study of auctions makes the most extensive use of models of individual behaviour based explicitly on the message requirements of the different institutions. This literature provides test comparisons of predicted behaviour, $m^i = \beta(E^i|I)$, with observations on individual choice, $\widehat{m}_i = \beta\left(\widehat{E}_i|I\right)$ for given realizations, $\hat{E}_i$ (such as values, $\widehat{v}_i$, where they are assigned at random). The large literature on experimental double auctions makes no such individual comparisons, because the theoretical literature had not yielded tractable models of individual bid-offer behaviour (but recent contributions by Friedman (1984), and Wilson (1984) are providing such models). Here as in most other areas of experimental research the comparisons are between the

predicted price–quantity outcomes of static theory (such as competitive, monopoly, and Cournot models), and observed outcomes. But double auctions have been studied (see references in Smith 1982) in a variety of environments; for example, the effect of price floors and ceilings have been examined (see references in Plott 1982). In all cases these studies are making comparisons. In *nomotheoretical* experiments one compares theory and observation, whereas in *nomoempirical* experiments one compares the effect of different institutions and/or environments as a means of documenting replicable empirical 'laws' that may stimulate modelling energy in new directions. The idea that formal theory must precede meaningful observation does not account for most of the historical development of science. *Heuristic* or exploratory experiments that provide empirical probes of new topics and new experimental methods should not be discouraged.

In industrial organization, and antitrust economics, experimental methods have been applied to examine the effects of monopoly, conspiracy, and alleged anticompetitive practices, and to study the concept of natural monopoly and its relation to scale economics, entry cost and the contestable markets hypothesis (see references in Plott 1982; Smith 1982; Coursey et al. 1984).

An important development in the experimental study of allocation processes has been the extension of experimental market methods to majority rule (and other) committee processes, and to market-like group processes for the provision of goods which have public or common outcome characteristics (loosely, public goods). These studies have examined public good allocation under majority (and Roberts') rules for committee including the effect of the agenda (see the references to Fiorina and Plott, and Levine and Plott in Smith 1982), and under compensated unanimity processes suggested by theorists (see the references in Coursey and Smith 1985). Generally, this literature reports substantial experimental support for the theory of majority rule outcomes, the theory of agenda processes (the sequencing of issues for voting decisions), and for incentive compatible models of the provision of public goods.

See Also

- ▶ Allais Paradox
- ▶ Efficient Allocation
- ▶ Preference Reversals
- ▶ Psychology and Economics

Bibliography

Battalio, R., J. Kagel, R. Winkler, E. Fisher, R. Basmann, and L. Krasner. 1973. A test of consumer demand theory using observations of individual consumer purchases. *Western Economic Journal* 11(4): 411–428.

Chamberlin, E. 1948. An experimental imperfect market. *Journal of Political Economy* 56: 95–108.

Coursey, D., and V. Smith. 1985. Experimental tests of an allocation mechanism for private, public or externality goods. *Scandinavian Journal of Economics* 86(4): 468–484.

Coursey, D., M. Isaac, M. Luke, and V. Smith. 1984. Market contestability in the presence of sunk (entry) costs. *Rand Journal of Economics* 15(1): 69–84.

Friedman, J. 1963. Individual behavior in oligopolistic markets: an experimental study. *Yale Economic Essays* 3(2): 359–417.

Friedman, D. 1984. On the efficiency of experimental double auction markets. *American Economic Review* 74(1): 60–72.

Hoggatt, A. 1959. An experimental business game. *Behavioral Science* 4(3): 192–203.

Kagel, J., R. Battalio, H. Rachlin, L. Green, R. Basmann, and W. Klemm. 1975. Experimental studies of consumer behavior using laboratory animals. *Economic Inquiry* 13(1): 22–38.

Ketcham, J., V. Smith, and A. Williams. 1984. A comparison of posted-offer and double-auction pricing institutions. *Review of Economic Studies* 51(4): 595–614.

Lakatos, I. 1978. In *The methodology of scientific research programmes, philosophical papers*, vol. 1, ed. J. Worrall and G. Currie. Cambridge: Cambridge University Press.

Mosteller, F., and P. Nogee. 1951. An experimental measurement of utility. *Journal of Political Economy* 59: 371–404.

Plott, C. 1982. Industrial organization theory and experimental economics. *Journal of Economic Literature* 20(4): 1485–1527.

Sauermann, H., and R. Selten. 1959. Ein Oligopolexperiment. *Zeitschrift für die Gesamte Staatswissenschaft* 115(3): 427–471.

Shubik, M. 1962. Some experimental non zero sum games with lack of information about the rules. *Management Science* 81(2): 215–234.

Siegel, S., and L. Fouraker. 1960. *Bargaining and group decision making*. New York: McGraw-Hill.

Smith, V. 1962. An experimental study of competitive market behavior. *Journal of Political Economy* 70: 111–137.
Smith, V. 1982. Microeconomic systems as experimental science. *American Economic Review* 72(5): 923–955.
Smith, V., A. Williams, K. Bratton, and M. Vannoni. 1982. Competitive market institutions: Double auctions versus sealed bid-offer auctions. *American Economic Review* 72(1): 58–77.
Williams, A., and V. Smith. 1984. Cyclical double-auction markets with and without speculators. *Journal of Business* 57(1) Pt 1: 1–33.
Wilson, R. 1984. Multilateral exchange. Working Paper No. 7, Stanford University, August.

Experimental Methods in Environmental Economics

Jason F. Shogren

Abstract

Experimental methods have long played a role in environmental economics. The strong link emerged due to the need to make decisions within the complex confluences of markets, missing markets, and no markets. Two broad areas of experimental work are discussed, institutional and valuation. Institutional experiments help reveal how good ideas for environmental protection can go badly with poorly understood rules and incentives; valuation experiments help illustrate how values for environmental protection depend on the socialization created, directly or indirectly, by the exchange institutions in operation.

Keywords

Asymmetric information; Bargaining; Coase Theorem; Common property resources; Endowment effect; Environmental economics; Experimental methods in environmental economics; Hypothetical bias; Institutional experiments; Land conservation; Non-market valuation; Pigouvian taxes; Pollution permits; Provision point mechanism; Public goods; Social costs; Transaction costs; Valuation experiments; Value elicitation; Willingness to accept compensation; Willingness to pay

JEL Classifications

C9

Environmental policy is designed within the confluence of markets, missing markets, and no markets. Within this mixture, economists offer working rules to help make outcomes more efficient, usually based on ideas formed by rational choice theory. The rules ask decision-makers to compare benefits in relation to costs, to account for the risks and gains across time and space for winners and losers, to facilitate the movement of resources from low-value uses to high-value uses, and to equate incremental gains per cost across policy actions. The environmental economic challenge is to find effective decision rules that will help move an economy towards efficient resource allocation in the face of market failure, for example, externalities, non-rival consumption, non-excludable net benefits, nonconvexities and asymmetric information (see Hanley, Shogren and White 2007).

Experimental methods have proven to be a useful tool in addressing this challenge. Environmental economists used experimental methods relatively early on, following the lead of Vernon Smith, Charles Plott and other pioneers. Experimental methods began to take hold in the 1980s, primarily in the area of non-market valuation (see Bohm 1972; Bennett 1983; Knetsch and Sinden 1984; Coursey et al. 1987). Today, experimental economic research is commonplace in environmental economic discussions and research programmes, with data being generated both in the laboratory and field (see for example the research in Cherry et al. 2007). Experiments in this area can be grouped broadly into two categories, *institutional* and *valuation*. Institutional experiments test-bed new institutions such as marketable pollution permits and ambient non-point pollution taxes prior to implementation; valuation experiments use the laboratory or field to study how people value goods and services that are not otherwise bought and sold in markets.

Institutional experiments build on traditional designs to test the efficiency of alternative exchange mechanisms under different economic circumstances. Usually the institutions under examination are those theoretically argued to correct for some market failure. Benefits and costs in these institutional experiments are *induced* by the experimenter – buyers have pre-assigned resale values; sellers have designated induced costs; and the goal is to measure the efficiency of a set of alternative incentive schemes.

In contrast, valuation experiments flip the institutional experiment on its head, using experimental methods to elicit preferences for some particular private or public good given alternative market and non-market circumstances. Here eliciting *homegrown* preferences or values – those residing within the minds of people – is of ultimate interest. Research in value elicitation has been environmental economics' most unique contribution to experimental economics. The work has produced insight into how the framing of a question affects values, how different demand-revealing incentives elicit different values, and how unintentional cues affect a person's value for a good. Consider now a few examples of institutional and valuation experiments used in environmental economics.

Institutional Experiments

Institutional experiments focus on evaluating market and non-market solutions to environmental problems. The key to these institutional experiments rests in the dialogue between the laboratory and potential or actual applications to environmental policy. For decades, environmental policy around the globe has been proposed and implemented in the real world with minimal input from insight gathered using experimental economics methods. Today, however, this is changing. Researchers are now using experiments to help understand and affect policy development, and this link between the laboratory and policy is probably more rigorously explored in environmental economics than any other area (Bohm 2003).

Institutional environmental economic experiments can be categorized as three broad areas – institutions to provide incentives to control externality problems that arise from pollution or land use; institutions to increase the voluntary provision of public goods, such as climate change, or to manage effectively common property, such as fishing zones; and institutions designed to manage resources through negotiation and cooperation, that is, the Coase theorem. We now briefly consider each in turn, starting with early work, moving to current applications, and general principles.

First, experiments examining economic solutions to externality problems began in earnest with Plott's (1983) work on Pigovian taxation. Plott designed a competitive market of buyers and sellers who trade a valuable good. After first establishing that traders ignored negative social costs in a competitive market, he explored whether Pigovian taxes or tradable permits could equate private incentives with social costs. Both increased efficiency with repeated trading periods and quickly hit 100% efficiency. Since then there has been an explosion of work examining incentive systems in a variety of settings, producing a growing and positive dialogue between policy proposals and insight from experimental studies.

Probably the most active area today remains the experimental work that tests the efficiency of tradable permit systems. Experimental methods have evaluated the efficacy of different trading rules in a variety of settings (for example, Bohm and Carlén 1999). An important early example is the US Environmental Protection Agency's Acid Rain emission trading. This work revealed a basic flaw in the original design of the permit auction run by the Environmental Protection Agency (EPA) (see Cason 1995; Cason and Plott 1996). The laboratory results revealed how the EPA could increase the efficiency of the auction by changing how permits were allocated. Originally, buyers and sellers submitted bids and offers for emission permits, and the EPA set the market price discriminatively off the demand curve by first matching the seller with the lowest offer to the buyer with the highest bid. The matching then

continued with the second lowest offer to the second highest bid, and so on, until the equilibrium quantity is reached. Rational sellers should see through this auction, and begin capturing rents by understating their true offer so they would be matched with a high bidder. Cason's laboratory results confirmed this intuition – sellers undercut each other to get into the high end of the market. The end result was an inefficient auction. Such lessons can be profitable, but insight like this should be made available before the regulatory tool is already in place, thus avoiding wasting resources due to inefficient design features. (For another important example comparing alternative trading institutions, see the tests of the RECLAIM market for the Los Angeles Basin by Ishikida et al. 2000).

Land conservation is a second area in which experimentally informed market designs have improved policy implementation. The Bush Tender auctions were designed to conserve land in Australia by creating a market where landowners bid to set aside specific units of land. Cason and Gangadharan (2004) examine how information about environmental benefits and a market clearing auction mechanism affect efficiency. Their results reveal an interesting pattern: people who did not know the environmental benefits provided by their private land were less likely to bid strategically in a conservation auction. Private ignorance reduces public expenditures. Based on this they suggest a provocative policy – a regulator might restrict the biological information publicly provided to landowners prior to running the auction. Another example of test-bedding is Parkhurst et al. (2002) agglomeration-bonus and smart-subsidy coordination game experiments, which illustrate an incentive scheme that can induce private landowners to create contiguous protected areas voluntarily. They compare a *smart subsidy* proposal, which creates an explicit link between neighbouring landowners with adjacent parcels, in relation to two standard policy options, compulsion and a standard fixed-fee subsidy. Their results show that a no-bonus mechanism always created fragmented habitat, whereas with the bonus, players found the first-best habitat reserve.

Second, environmental policy has long confronted the inherent efficiency issues associated with public goods and common property resources. These experimental games capture the elemental economic problem that drives many environmental goods: non-rival and non-exclusive consumption lead to free riding and inefficient production levels. Experimental evidence reveals neither complete free riding nor full cooperation (Ledyard 1995). As noted by Ostrom (2000), three types of people commonly inhabit public-good and common-property experiments: the standard rational egoists, the conditional cooperators, and the willing punishers. Conditional cooperators cooperate when they expect others to reciprocate; otherwise they do not. Within the standard game (see public goods experiments) rules can be manipulated to induce more or less cooperation depending on the mix of subject types, marginal payoffs, group size, communication, and voting with third-party enforcement.

For global environmental goods like climate change, a key policy issue is the impact on efficiency when a collective agreement has costly third-party enforcement. Punishment of free riders is a second-order public good; cooperation means bearing some private cost to sanction others. One relevant policy question is whether an institution based on a voting rule with a punishment mechanism can work to increase contributions to a public good. Kroll, Cherry and Shogren (2007) examined this in the laboratory and observed that voting alone does not increase cooperation; rather, if voters can pay to punish violators, contributions increase significantly. Overall efficiency for a voting-with-punishment rule exceeds the level observed for a voting-without-punishment rule. This result has implications for how policymakers think about institutions such as International Environmental Agreements (IEA), which are more likely to be successful if one nation is willing to act like the 'global police', and pay the costs of punishing violators (Barrett 2003).

Another real-world policy issue is whether policymakers can use economic incentive devices in real-world applications to reveal public good

demand (Bohm 1972). Any proposed system has to 'work' – to provide the good when benefits exceed the provision costs – and has to be straightforward enough to be implemented in the field, characteristics that can be tested in the laboratory. One such mechanism is the *provision point* mechanism: if contributions meet or exceed a targeted provision cost, the public good is supplied to the group; otherwise, it is not. In a design that mimics field conditions, Rondeau, Schulze and Poe (1999) explored a mechanism in which contributions are returned if costs were not met. Their results suggest the provision point mechanism was 'demand revealing in aggregate' for a large group with heterogeneous preferences, suggesting that a relatively simple mechanism could be used in the field to elicit preferences, leading to the efficient provision of a public good.

Third, many observers and policymakers see place-based collaboration and bargaining as the future of environmental policy – arguing for more local control through negotiation and accountability (for example, Sabel et al. 2000). Collaborative decision-making groups have begun to flourish in rural settings such as the western United States, and now number in the hundreds, ranging from informal grass-roots gatherings to government-mandated advisory councils. To an economist, this is the direct application of the Coase theorem – parties in dispute negotiating on a jointly acceptable agreement over resource use. Starting with Hoffman and Spitzer (1982), researchers have used experimental methods to test the robustness of collaborative decision-making underlying the Coase theorem. Hoffman and Spitzer's initial results supported the Coase theorem in that bargains were highly efficient. Harrison and McKee (1985) confirmed that Coasean bargaining under unilateral and joint property rights regimes can be efficient. Both experiments assumed the transaction costs of bargaining were zero. Recent experimental work has explored how bargaining efficiency is affected by positive transaction costs and addition friction due to large numbers of bargainers, property right insecurity, delay costs, imperfect contract enforcement, asymmetric information, and uncertain final authority. The lesson from over two decades of Coase bargaining research is that people have to address the nature of transaction costs and friction.

We illustrate using Rhoads and Shogren's (2003) policy-driven Coasean bargaining experiment. Experts see two elements of consensus-based environmental protection as crucial for effective regulatory outcomes: *final authority*, so that a collaborative agreement is binding; and *information symmetry*, in which bargainers create a common information pool about player payoffs. Their results are consistent with the findings of the experts: final authority and information symmetry were necessary conditions for efficient Coasean bargaining. Without final authority, efficiency falls by two-thirds, and falls further with asymmetric information. If the policy objective is to make a negotiated agreement efficient, the policy challenge is to understand the trade-offs associated with granting or denying final authority to the local bargainers.

In summary, the general principle in institutional environmental economics experiments that has emerged over the years is that the germ of a good idea can be codified into a bad one if the rules of implementation trigger unintended incentives that undercut the efficiency of the system. Experimental methods can be used to reveal which good ideas are actually beneficial to control externalities, provide public goods, and facilitate collaboration – and which ideas are ultimately counterproductive.

Valuation Experiments

Economists also use experimental methods to understand better the behavioural underpinnings of environmental valuation. Experiments can be used to address incentive and contextual questions that arise in assessing values through direct statements of preferences. Three general areas have emerged: rational valuation, direct elicitation of values, and exploring the effectiveness of hypothetical non-market valuation surveys (see Shogren 2006).

First, economists assume people can provide rational statements of their preferences and values

towards the environment. Rather than assume that people make rational choices and reveal consistent values for environmental protection, environmental economists use experiments to examine whether people's choices and stated values meet these criteria. Enough evidence of behavioural anomalies now exists to undercut this presumption (Kahneman and Tversky 2000). Without an exchange institution to arbitrage his or her irrational choices, the unsocialized person can engage in behaviours inconsistent with rational choice theory (see Akerlof 1997).

The key behavioural regularity that potentially undercuts all valuation work is the WTP–WTA gap. Rational choice theory suggests that with small income effects and many available substitutes, the willingness to pay (WTP) for a commodity and the willingness to accept (WTA) compensation to sell the same commodity should be about equal. But evidence suggests that WTA exceeds WTP by up to tenfold. The experimental WTP–WTA work can be divided into two camps: research that suggests the gap is based on a psychological *endowment effect* and that which points to weak market institutions. A person who assigns greater value to a good he or she already owns exhibits the endowment effect, which leads to higher WTA to sell the good than WTP to buy the identical good (Kahneman et al. 1990). The market experience explanation says that people have naive expectations about what they can sell the good for outside an active market place. This experimental work showed that market-like experience can remove the gap (see Shogren et al. 1994, 2001).

Second, valuation experiments are used to measure actual values for public and private goods (Lusk and Shogren 2007). Direct valuation experiments are designed so that people buy and sell actual goods to elicit real values, in which researchers test how alternative exchange institutions affect these values. They entail real payments and binding budget constraints, and use auctions to sell goods for money, albeit within a stylized setting. Experimental designs are used to understand the balance between laboratory control and natural context, enabling researchers to learn things about behaviour that would have been impossible to discover from alternative tools. Subtle changes in experimental procedure affect behaviour, such as paying people before as opposed to after bidding, reporting the market-clearing price, and the novelty of the good.

Third, in the 1980s, Coursey and Schulze (1986) hoped the laboratory would be used more to test-bed field surveys. Today, in 2007, experiments are commonly employed to address problems in stated preference surveys such as hypothetical bias, calibration, surrogate bidding, and incentive compatibility. For instance, experiments have revealed time and again that *hypothetical bias* is real – people frequently promise more than they actually deliver. Experimental work has focused on trying to measure the degree of bias and what methods can be used to eliminate or reduce it in survey work. A good example is Cummings and Taylor (1999) who find that they can remove the hypothetical bias by telling a respondent about it.

In summary, choices and economic values emerge in the social context of an active exchange institution, and thus the measurement of value should not be separated from the interactive experience provided by an exchange institution. Institutions and the institutional context matter because experience can make rational choice more transparent to a person. Institutions also dictate the rules under which exchange occurs, and these rules can differ across settings. People can interpret differently the information conveyed by such settings. The reality is that most people make allocation decisions in several institutional settings each day – markets, missing markets, and unidentified markets. How does this institutional mix affect how people make their choices and form or state their preferences for environmental protection? This question is fundamental because it gives a reason for the purposeful actions underlying all valuation work.

Experimental work like the *rationality spillover* treatments in Cherry, Crocker and Shogren (2003) reveal that exposure to competition and discipline is needed to achieve rationality. In becoming rational, people refine their statements of value to better match their preferences. The contact with others who are making similar

decisions in an exchange institution puts in context the economic maxim that choices have consequences and stated values have meaning for environmental valuation. Relying on rational theory to guide environmental valuation and policy makes more sense if people make, or act as if they make, consistent and systematic choices about certain and risky events. Valuation work in the laboratory needs to continue to address the economic conditions under which the presumption of rationality is supported and when it is not, which in turn has implications for the values we directly elicit.

Concluding Remarks

Through the use of experimental methods, environmental economists now understand better how people learn about and react to incentives, institutions and information. They can compare how decisions are made with and without real economic commitments, within and without active exchange institutions, and with and without signals of value. They can then delve into what the results suggest for *ex ante* questionnaire design, *ex post* statistical evaluation, and, more importantly perhaps, economic theory itself. The environmental economics literature continues to follow the classic experimental strategy: start simply and add complexity slowly so as to understand which factors matter, and why.

In addition, all experiments in environmental economics reveal the perpetual tension between *control* and *context*. At the core, the experimental method is about *control*. One controls the experimental circumstances by trying to change only one variable at a time, which will reduce problems of confounding. Without control, it is unclear whether unpredicted behaviour is due to a poor theory or experimental design, or both. In contrast, others argue that *context* is desirable to avoid a setting that is too sterile and too removed from reality for something so real as environmental policy. Context affects participants' motivation.

Finally, as evidence continues to accumulate, a clearer and more definitive picture will emerge of how our institutions affect the efficiency and perceived value of environmental policies. The future of experimental work will be to design institutions that address the combination of market failure and behavioural anomalies. Otherwise we could find environmental economics falling into a new second-best problem: if we correct market failure without addressing behavioural biases, we might actually reduce overall social welfare.

See Also

- ▶ Coase Theorem
- ▶ Environmental Economics
- ▶ Experimental Economics
- ▶ Market Failure
- ▶ Pollution Permits
- ▶ Public Goods Experiments
- ▶ Value Elicitation

Bibliography

Akerlof, G. 1997. Social distance and social decisions. *Econometrica* 65: 1005–1027.

Barrett, S. 2003. *Environment and statecraft: The strategy of environmental treaty-making*. New York: Oxford University Press.

Bennett, J. 1983. Validating revealed preferences. *Economic Analysis and Policy* 13: 2–17.

Bohm, P. 1972. Estimating demand for public goods: An experiment. *European Economic Review* 3: 111–130.

Bohm, P. 2003. Experimental evaluations of policy instruments. In *Handbook of environmental economics*, ed. K.G. Mäler and J. Vincent, vol. 1, 438–460. Amsterdam: North-Holland.

Bohm, P., and B. Carlén. 1999. Emission quota trade among the few: Laboratory evidence of joint implementation among committed countries. *Resource and Energy Economics* 21: 43–66.

Cason, T. 1995. An experimental investigation of the seller incentives in the EPA's emission trading auction. *American Economic Review* 85: 905–922.

Cason, T., and C. Plott. 1996. EPA's new emissions trading mechanism: A laboratory evaluation. *Journal of Environmental Economics and Management* 30: 133–160.

Cason, T., and L. Gangadharan. 2004. Auction design for voluntary conservation programs. *American Journal of Agricultural Economics* 86: 1211–1217.

Cherry, T., T. Crocker, and J. Shogren. 2003. Rationality spillovers. *Journal of Environmental Economics and Management* 45: 63–84.

Cherry, T., S. Kroll, and J. Shogren. 2007. *Experimental methods, environmental economics*. London: Routledge.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.
Coursey, D., and W. Schulze. 1986. The application of laboratory experimental economics to the contingent valuation of public goods. *Public Choice* 49: 47–68.
Coursey, D., J. Hovis, and W. Schulze. 1987. The disparity between willingness to accept and willingness to pay measures of value. *Quarterly Journal of Economics* 102: 679–690.
Cummings, R., and L. Taylor. 1999. Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review* 83: 649–665.
Hanley, N., J. Shogren, and B. White. 2007. *Environmental economics in theory and practice*. 2nd ed. London/New York: Palgrave.
Harrison, G., and M. McKee. 1985. Experimental evaluation of the Coase theorem. *Journal of Law and Economics* 28: 653–670.
Hoffman, E., and M. Spitzer. 1982. The Coase theorem: Some experimental tests. *Journal of Law and Economics* 25: 73–98.
Ishikida, T., J. Ledyard, M. Olson, and D. Porter. 2000. Experimental testbedding of a pollution trading system: Southern California's RECLAIM emissions market. In *Research in experimental economics*, ed. R.M. Isaac, vol. 8. Greenwich: JAI Press/, Elsevier Science.
Kahneman, D., J. Knetsch, and R. Thaler. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98: 1325–1348.
Kahneman, D., and A. Tversky. 2000. *Choices, values and frames*. Cambridge: Cambridge University Press.
Knetsch, J., and J.A. Sinden. 1984. Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of values. *Quarterly Journal of Economics* 99: 507–521.
Kroll, S., T. Cherry, and J. Shogren. 2007. Voting, punishment and public goods. *Economic Inquiry* 45: 557–570.
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.
Lusk, J., and J. Shogren. 2007. *Experimental auctions: Methods and applications in economic and marketing research*. New York: Cambridge University Press.
Ostrom, E. 2000. Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14(3): 137–148.
Parkhurst, G., J. Shogren, C. Bastian, P. Kivi, J. Donner, and R. Smith. 2002. Agglomeration bonus: An incentive mechanism to reunite fragmented habitat for biodiversity conservation. *Ecological Economics* 41: 305–328.
Plott, C. 1983. Externalities and corrective policies in experimental markets. *Economic Journal* 93: 106–127.
Rhoads, T., and J. Shogren. 2003. Regulation through collaboration: Final authority and information symmetry in environmental Coasean bargaining. *Journal of Regulatory Economics* 24: 63–89.
Rondeau, D., W. Schulze, and G. Poe. 1999. Voluntary revelation of the demand for public goods using a provision point mechanism. *Journal of Public Economics* 72: 455–470.
Sabel, C., A. Fung, and B. Karkkainen. 2000. *Beyond backyard environmentalism*. Boston: Beacon Press.
Shogren, J. 2006. Experimental methods and valuation. In *Handbook of environmental economics*, ed. K.G. Mäler and J. Vincent, vol. 2. Amsterdam: North-Holland.
Shogren, J., S. Shin, D. Hayes, and J. Kliebenstein. 1994. Resolving differences in willingness to pay and willingness to accept. *American Economic Review* 84: 255–270.
Shogren, J., S. Cho, C. Koo, J. List, C. Park, P. Polo, and R. Wilhelmi. 2001. Auction mechanisms and the measurement of WTP and WTA. *Resource and Energy Economics* 23: 97–109.

Experiments and Econometrics

Daniel E. Houser

Abstract

'Experimetrics' refers to formal procedures used in designed investigations of economic hypotheses. Fundamental experimetric contributions by Ronald A. Fisher provided the foundation for a rich literature informing the design and analysis of economics experiments. Key components of this foundation include the concepts of randomization, independence and blocking. Experimetric analysis plays a central role in advancing economic models, and will gain further importance as scholars adopt increasingly sophisticated designed research programmes to illuminate positive economic theory.

Keywords

Blocking; Causal inference; Exact Test (R. A. Fisher); Experimental economics; Experiments and econometrics; Experimetrics; Fisher, R. A.; Individual learning in games; Neuroeconomics; Quantal response equilibrium; Randomization

JEL Classifications
C9

Introduction

'Experimetrics' refers to formal procedures used in designed investigations of economic hypotheses. A series of pathbreaking experimetric contributions by Ronald A. Fisher, written largely during the 1920s and early 1930s, elucidated fundamental concepts in the design and analysis of experiments (see, for example, Box 1980, for a survey). He was first to obtain rigorous experimetric results on the importance of randomization, independence and blocking, and he created many powerful analysis tools that remain widely used, including Fisher's nonparametric Exact Test (Fisher 1926; see also Fisher 1935).

Controlled experiments allow compelling scientific inferences with respect to hypotheses of interest. Many economic experiments inform hypotheses regarding primitives assumed to be constant within an experiment (for example, preferences or decision strategies), or the effects on economic outcomes of changes in institutions (for example, comparing different auction rules or unemployment regulations; see experimental economics). One conducts controlled experiments to inform economic hypotheses because relevant naturally occurring data typically include noise of unknown form and magnitude outside the investigator's control. Econometric procedures can go some distance towards solving this problem, but even sophisticated approaches often allow only limited conclusions.

For example, suppose one wanted to investigate the (causal) effect of caffeine on heart rhythms. One approach is to obtain a random sample of 'heavy' coffee drinkers and compare them with a random sample of people who do not use caffeine. Because it is not possible with naturally occurring data to control the reason a person falls into a category, discovering that people with greater caffeine consumption have more cardiac episodes need not imply a causal caffeine effect. The reason is that a preference for coffee may stem from a biological characteristic that is itself causally tied to irregular cardiac events.

An advantage of designed investigations is that they allow cogent inference regarding causal effects through the appropriate use of randomization, independence and blocking.

Randomization

Experiments with randomized designs allow compelling causal inference. The reason is that randomly assigning participants to treatments, and randomly assigning treatments to dates and times, minimizes the possibility of systematic error. In the caffeine example, intentionally assigning heavy caffeine drinkers exclusively to a caffeine treatment generates a systematic error and invalidates causal inference. However, an experiment where subjects are randomly assigned to caffeine and no-caffeine treatments independent of their typical caffeine use allows one to draw appropriate inferences regarding causal relationships.

Independence

Randomization also helps to ensure independence both within and between treatments' observations. Loosely speaking, observations are independent if information about one observation does not provide information about another. Independence is critical for many experimetric analyses, and its failure can lead to misleading conclusions. An objective randomization procedure for treatment assignments insures against the possibility that participants in one treatment might unintentionally systematically vary from other treatments' participants.

Blocking

Causal relationships can be assessed with greater precision through 'blocking'. Blocking is a design procedure with which an experimenter can separate treatment effects from nuisance sources of data variation. In the above, heart rhythms might be affected by both caffeine and anxiety over the process of measuring heart rhythms. Especially because it is expected to

differ between participants, anxiety is a source of nuisance variation that clouds inferences regarding caffeine effects. To address this one could 'block' by participant. This involves measuring each subject both with and without caffeine (in separate, randomly ordered trials). Caffeine effects are measured as the difference between trials, thus mitigating noise due to individual anxiety effects.

Experimetrics Toolbox

Although many specialized experimetric tools have been developed, the experimetrics toolbox also includes a large number of general purpose procedures that have become standard in the experimental economics literature. A regular concern is that independence is not satisfied. The failure of independence can occur because of 'session' effects, meaning that there is less behavioural variation within than between sessions. Violations of independence can also occur if repeated measurements are taken on the same individual due to individual effects. Standard procedures can address this. Sessions can be treated as fixed effects, and random effects can be used to control for individual differences. The resulting 'mixed effect' model can be analysed using standard parametric, panel-data procedures (see, for example, Frechette 2005).

Also in the toolbox is the McKelvey and Palfrey (1995) 'quantal response equilibrium' (QRE) framework (see quantal response equilibria). QRE is a parametric procedure for analysing data from finite games. The key idea is to incorporate errors into players' best response functions, thus creating 'quantal response' functions. This results in an extremely flexible model that can rationalize a wide variety of behaviours. Haile et al. (2006) point out that this flexibility comes at a cost: in general QRE can rationalize any distribution of behaviour in any normal form game, and imposes no falsifiable restrictions without additional assumptions on the stochastic components of the model. Thus, those who wish to implement QRE analyses face the experimetric challenge of creating designs within which such assumptions are defensible.

For reasons including sample size and robustness, the experimetrics toolbox includes many nonparametric procedures (see Siegel and Castellan 1988, for a user- friendly textbook treatment of popular nonparametric approaches). For example, Mann–Whitney tests, and their k-sample generalization due to Jonckheere (1954), are frequently used to compare medians among treatments' data. Also common is Fisher's Exact Test, which uses all the information in the data and is the most powerful nonparametric approach to inference with respect to differences among treatments. Its use is limited by the fact that it can be computationally cumbersome to implement when the numbers of treatments or observations are large.

External Validity

An experiment's conclusions are 'externally valid' if they can be extrapolated to other environments. To rigorously address external validity requires that the source of treatment effects can be identified, which in turn implies a fundamental rule of experiment design: within any good experiment, any treatment can be matched with another that differs from it in exactly one way.

External validity is both important and subtle. For example, consider the well- known 'dictator game' where one participant is assigned the role of 'dictator', and the other 'receiver'. The dictator is given $20, and the receiver nothing. The dictator is told to split the $20 between herself and her receiver in any way she likes, after which the experiment ends. A widely replicated result is that a large fraction of dictators send half ($10) to an anonymous stranger, and one might question whether this finding is externally valid. In particular, there is no evidence that this behaviour is prevalent among winners of naturally occurring lotteries.

There are clear similarities between the situations of lottery winners and dictators. Still, the fact that actions of dictators in laboratory games do not match actions of lottery winners does not

necessarily mean that dictator games lack external validity. The reason is that identical decision strategies can imply different decisions in different environments. For example, recent research provides compelling evidence that dictators' decisions are tightly connected to their beliefs regarding the decisions of others who have faced this same situation: dictators give because they believe other dictators give (Bicchieri and Xiao 2007). This mechanism plausibly guides decisions in naturally occurring environments. In particular, lottery winners do not give because they believe other lottery winners do not give large fractions of their winnings to anonymous strangers. Thus, external validity does not require that one be able to match actions in an experiment to actions in another environment. Rather, an experiment is externally valid if one can extrapolate to novel contexts its conclusions with respect to individual or strategic decision processes.

Applied Experimetrics Research

An important application of experimetrics is to discriminate between many competing theories of learning that have emerged (see individual learning in games). Doing this includes significant experimetric challenges, as it requires one to account for heterogeneity in the way subjects learn. The reason is that not doing so will tend to bias fit statistics in favour of reinforcement (and hybrid) models. Wilcox (2006) shows the reason is that reinforcement models condition behaviour on informative functions of past choices, and in the presence of learning heterogeneity these choices will carry idiosyncratic parameter information not otherwise incorporated into the specification. Having said this, it is also the case that many data-sets from typical learning experiments can be roughly equally well described by many different learning models (Salmon 2001). Consequently, the 'best' model can be highly sensitive to the particular criterion one uses for model selection, as well as the particular experiment under consideration (Feltovich 2000). As a result, in-sample fit is often good, but this does not necessarily imply that much has been learned about the way in which people actually learn and make choices (Salmon 2001).

Knowing how people make choices is critical to advance both economic theory and institution design. Consequently, a significant experimetric literature explores how people make decisions in complex environments, with a focus on characterizing the nature and number of different 'decision rules' at use in a population. Most approaches to accomplishing this require pre-specifying the decision rules the researcher believes people could follow, and then using choice data to assign one of those rules to each member of the population (see, for example, El-Gamal and Grether 1995). However, in some cases one might be unwilling or unable to pre-specify the decision rules, and it turns out that doing so is not necessary. In particular, Houser et al. (2004) detail a Bayesian experimetric procedure that uses individual choice data to determine endogenously the nature and number of decision rules in a population. The approach requires only that one specify the information relevant to individuals' decisions.

Substantive experimetric advances have been obtained in far too many areas to detail here. Although no general survey is available, Houser et al. (2004), Ashley et al. (2005), and Loomes (2005), include excellent summaries of experimetric contributions to a variety of widely-studied games and decision problems.

Conclusion

Experimetrics continues to evolve as scholars adopt highly sophisticated design and analysis procedures to inform new questions. A ready example is the rapidly expanding research in neuroeconomics (see neuroeconomics). The massive spatial- panel data structure that characterizes brain images poses unique inferential problems. Progress on these problems requires significant complementary innovations to both design and analysis strategies. The resulting experimetric advances are sure to have significant impact on economic theory and policy analysis.

See Also

- ▶ Experimental Economics
- ▶ Experimental Economics, History of
- ▶ Experimental Methods in Economics
- ▶ Individual Learning in Games
- ▶ Neuroeconomics
- ▶ Quantal Response Equilibria
- ▶ Smith, Vernon (Born 1927)

Bibliography

Ashley, R., S. Ball, and C. Eckel. 2005. *Motives for giving. A reanalysis of two classic public goods experiments*. Manuscript, Virginia Institute of Technology.

Bicchieri, C., and E. Xiao. 2007. *Do the right thing: But only if others do*. Manuscript, University of Pennsylvania.

Box, J.F. 1980. R.A. Fisher and the design of experiments, 1922–1926. *American Statistician* 34: 1–7.

El-Gamal, M.A., and D.M. Grether. 1995. Are people Bayesian? Uncovering behavioral strategies. *Journal of the American Statistical Association* 90: 1137–1145.

Feltovich, N. 2000. Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. *Econometrica* 68: 605–641.

Fisher, R.A. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 33: 503–513.

Fisher, R.A. 1935. *Design of experiments*. Edinburgh: Oliver and Boyd.

Frechette, G. 2005. *Session effects in the laboratory*. Manuscript, New York University.

Haile, P.A., A. Hortacsu, and G. Kosenok. 2006. *On the empirical content of quantal response equilibrium*. Mimeo, Yale University.

Houser, D., M. Keane, and K. McCabe. 2004. Behavior in a dynamic decision problem: An analysis of experimental evidence using a Bayesian type classification algorithm. *Econometrica* 72: 781–822.

Jonckheere, A.R. 1954. A distribution-free k-sample test against ordered alternatives. *Biometrika* 41: 133–145.

Loomes, G. 2005. Modelling the stochastic component of behavior in experiments: Some issues for the interpretation of data. *Experimental Economics* 8: 301–323.

McKelvey, R.D., and T.R. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10: 6–38.

Salmon, T.C. 2001. An evaluation of econometric models of adaptive learning. *Econometrica* 69: 1597–1628.

Siegel, S., and N. Castellan Jr. 1988. *Nonparametric statistics for the behavioral sciences*. 2nd ed. Boston: McGraw Hill.

Wilcox, N. 2006. Theories of learning in games and heterogeneity bias. *Econometrica* 74: 1271–1292.

Explanation

Julian Reiss

Abstract

Explaining socio-economic phenomena is one important aim of economics. There is very little agreement, however, on what precisely constitutes an adequate economic explanation. Starting from the very influential but defective 'deductive-nomological model' of explanation, this article describes and criticizes the major contemporary competitors for such an account (the probabilistic–causal, the mechanistic–causal and the unificationist models) and argues that none of them can by itself capture all aspects of a good explanation. When seeking to explain a socio-economic phenomenon it should therefore be borne in mind that different types of explanation serve different purposes.

Keywords

Causality in economics and econometrics; Economic laws; Equilibrium; Explanation; Laws of nature; Methodology of economics; Positive economics; Probability; Statistical inference

JEL Classifications

B4

In the early 1950s Milton Friedman famously declared that the 'ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful (that is, not truistic) predictions about phenomena not yet observed' (Friedman 1953, p. 7). Today, after the demise of logical positivism in philosophy and positivistic trends in economics, economists tend to regard the explanation of phenomena as one legitimate aim of economics besides the more directly policy-oriented aims of prediction and control. Perhaps, following Friedman, explaining

a phenomenon is primarily of instrumental value for the preparation and guidance of policy. But perhaps economists seek to explain in order to increase our understanding of the economic world, for purely cognitive reasons. Whether derivative or fundamental, explanation is a major goal that economists pursue and understanding what exactly is sought is an important task for economic methodology.

An adequate account of explanation in economics should satisfy at least three desiderata:

(a) it should be *descriptively adequate*; that is, it should be consistent with economic practice;
(b) it should be *epistemically adequate*; that is, it should give reason to believe that that which it identifies as an explanation is indeed explanatory; and
(c) it should be *empirically adequate*; that is, it should not identify something as an explanation unless it is based on sufficient evidence.

The so-called deductive–nomological or DN model of explanation (Hempel and Oppenheim 1948) can rightly be regarded as the received view of scientific explanation. Although the theory is now generally regarded as untenable, it is it is useful to consider its guiding ideas as a starting point because its flaws motivate the alternative, more satisfactory accounts.

The Deductive–Nomological Model

According to the DN model, an explanation is an argument whose premises constitute the so-called explanans (or 'that which explains') and whose conclusion constitutes the so-called explanandum (or 'that which is to be explained'). The explanandum will usually be a description of a noteworthy singular event (such as 'Black Monday', 'the rise of the dot.com industry' or 'the collapse of the Tiger economies') or a repeated pattern of events, which may be called a 'phenomenon' (such as 'hyperinflations', 'the J-curve effect' or 'the price drop of cars that have just left the showroom').

The adjectives 'deductive' and 'nomological' indicate that the argument must meet at least two criteria in order to count as an explanation. First, the argument must be deductively valid, that is, the explanandum must follow logically from the explanans. Second, among the premises of the explanans there must be at least one law of nature (the Greek word *nómos* means habit or law). Typically, it is also demanded that the premises of the explanans be true or at least verified. However, none of these criteria is individually necessary nor are the criteria jointly sufficient.

In many cases explanations are probabilistic rather than deterministic and thus the explanandum does not always logically follow from the explanans. John Doe's exposure to asbestos explains his contraction of lung cancer but the statement 'John contracted lung cancer' is not entailed by the statement 'John was exposed to asbestos'. Second, and related, laws of nature in the sense of universal regularities are few and far between, especially in non-fundamental sciences such as economics. All so-called 'laws' in economics, such as the law of supply and demand, the iron law of wages, Okun's law, Say's Law and so forth are, at best, true *ceteris paribus*, that is, if nothing intervenes and relative to a specific institutional structure. For example, we can use the law of supply and demand to predict that demand for a good will decrease when a tax is imposed. However, depending on what else happens in the economy actual demand may or may not decrease. If disposable incomes rise sufficiently or if preferences change in the right way, demand may in fact increase.

Third, it is not clear whether laws in the sense used by proponents of the DN model are explanatory at all. Suppose that it is a law – a universal regularity – that economic expansions follow monetary expansions. Economists no doubt regard knowledge of this kind as very valuable, but unless more is told about the relationship it would hardly count as explanatory. The DN model is therefore neither descriptively nor epistemically adequate.

In response to these and other difficulties of the DN model (for a valuable discussion of many of the

criticisms, see van Fraassen 1980, pp. 103–29) philosophers have developed alternative accounts of explanation. One tradition holds that to explain a phenomenon means to cite the causes of the phenomenon. It therefore roughly agrees with the 'nomological' part of the DN model but replaces the notion of law with that of cause. Another tradition holds that to explain a phenomenon means to show how it fits into a systematization of our beliefs about the world. It agrees with the 'deductive' part of the DN model and insists that good explanations are those that unify diverse sets of beliefs. Both traditions can be found in economics, and both come in two variants.

The Probabilistic–Causal Model

The chief difficulty for the causalist, who maintains that to explain a phenomenon is to provide information about its causes, is to elucidate the notion of cause. We believe that a tightening of the money stock explains the subsequent increase in interest rates; a change in minimum wages explains changes in the employment rate; veteran status explains earnings. In none of these cases is there a universal regularity between event-types; rather, earlier events appear to be probabilistic causes in the sense that they are *statistically relevant*.

One view thus held that event X explains event Y if the probability of Y in some population described by Z is different when X is present from when it is absent: $P(Y|\ X,\ Z) \neq P(Y|Z)$ (cf. Salmon 1971). In econometrics this idea is akin to the notion of a multiple regression:

$$Y = \alpha X + \boldsymbol{\beta}\mathbf{Z} + \varepsilon,$$

where Y is the explained variable, X is the explanatory variable and $\mathbf{Z}$ is a vector of background variables. X is statistically relevant to Y if and only α is different from zero, and can thus be used to explain Y.

Not all statistically relevant events appear to be explanatorily relevant, however. A drop in the barometer reading raises the probability of a storm but the barometer reading does not explain the storm. It is a common cause – the change in atmospheric pressure – that explains its joint effects, the barometer reading and the storm.

This suggests that X plus the set of background factors must constitute the full set of causes of the explained variable (cf. Cartwright 1983, Essay 1): in any population that is homogenous with respect to atmospheric pressure, the barometer reading is statistically irrelevant to the occurrence of the storm. An obvious drawback of this account is that it asks for immense amounts of background knowledge for identifying explanatory factors from statistics. It requires that all other causes of a phenomenon (that is, all confounding factors) to be known or known to be distributed equally between a treatment and a control group, as in a randomized trial. Given the complexities of the social world, one can expect this requirement to be met only exceptionally.

There are also more principled difficulties. One problem arises because some factors may act differently depending on what other causes are present. An increase in the money stock may have different effects on the economy depending on the interest rate and investor behaviour. In extreme situations increasing money may have no effects on the economy at all, and so the government's ability to conduct monetary policy is thus incapacitated. It therefore seems exaggerated to demand that explanatory factors raise the probability of the explained variable in *all* causally homogeneous populations (which is presupposed by the linear models favoured by econometricians). But is it enough for a factor to raise the probability of the effect in one single population or should it raise its probability on average (cf. Dupré 1984)? Moreover, when factors act genuinely probabilistically, factors can be statistically relevant despite the fact that they are not explanatorily relevant, even when all other causes have been included. Suppose that in some causally homogeneous background Z, money M is a probabilistic cause of both nominal income Y as well as the level of prices L. Let $P(Y|M, Z) = P(L|M, Z) = 0.8$. Now, let money cause income on precisely those occasions that it causes prices and vice versa. Then, $1 = P(Y|L, M, Z) > P(Y|M, Z) = 0.8$ even though the change in prices does

not explain the change in nominal income – it is a mere correlate (cf. Cartwright 1999, ch. 5).

But it is important to keep practical and epistemic issues apart. *If* one knows that *C* is a probabilistic cause of a phenomenon of interest *E*, then there is no reason to deny that one can use *C* in an explanation of *E*. The epistemic adequacy of the probabilistic–causal model derives from the general acceptance of causes as explanatory factors. However, finding out if *C* is a probabilistic cause of *E* will often face insurmountable practical difficulties.

The Mechanistic–Causal Model

In philosophy of science, the mechanistic–causal model has been mostly associated with the name Wesley Salmon (see for instance Salmon 1984). It attempts to improve upon the probabilistic model on two counts. On the one hand, for practical purposes it may be easier to find out whether *C* causes *E* by investigating whether or not there is a mechanism running from *C* to *E* than by statistical inference. For instance, Milton Friedman and Anna Schwartz (1963, p. 59) write:

> However consistent may be the [statistical] relation between monetary change and economic change, and however strong the evidence for the autonomy of the monetary changes, we shall not be persuaded that the monetary changes are the source [that is, cause] of the economic changes unless we can specify in some detail the mechanism that connects the one with the other.

Indeed, if *C* causes *E* we expect there to be a mechanism running from *C* to *E*. Evidence about a mechanism from *C* to *E* can thus provide evidence for a causal connection. In turn, according to this view, the mechanism can be used to explain *E*.

On the other hand, causal explanations that are based on statistical inferences often cite relationships among aggregate factors such as the money stock, the unemployment rate, inflation and so on, and can arguably be said to be somewhat shallow. Perhaps a monetary expansion can be used to explain a subsequent economic expansion because there is statistical evidence that the former is the cause of the latter. In this way one learns at best *that* the monetary change causes the economic change. Describing the transmission mechanism one further learns *how* the monetary change causes the economic change. The explanation is thus arguably more detailed, deeper.

It cannot be said, however, that the mechanistic account wins unequivocally over the probabilistic account on both fronts. In order to meet the empirical adequacy desideratum, a mechanistic explanation must be based on evidence no less than a probabilistic-causal explanation. A mere 'sketch' of a mechanism (such as the sketch of the transmission mechanism that follows above quotation by Friedman and Schwartz) does not explain anything. Usually, mechanistic explanations cite relationships among individuals, their preferences and external constraints. The argument that such hypotheses are more readily verifiable (for instance, because they may be verifiable by introspection) goes back at least to the writings of John Stuart Mill (1830). But it is not clear whether it is always easier to provide evidence for causal mechanisms that run at the micro level than for aggregate causal relationships. For instance, the problem of confounding factors is in no way confined to statistical inferences among aggregate variables, and, at the micro level, can only seemingly be alleviated by assuming away the operation of confounders a priori.

Moreover, although with some justification it can be said that mechanistic explanations are deeper than aggregate explanations, there are situations in which information about exactly how some variable influences another is entirely irrelevant. A policymaker, for instance, may be more interested in what is common among expansion episodes rather than in the exact processes that made them happen – which may be different on each occasion.

How Models Explain: Unificationism

Let us now move from the applied side to the more theoretical side of economics. Consider the following quotation (Akerlof 1970):

> From time to time one hears either mention of or surprise at the large price difference between new

> cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a 'new' car. We offer a different explanation.

Akerlof then describes an asymmetric-information model in which low-quality second-hand cars drive higher-quality cars out of the market, which leads to a decrease in average quality and prices. One way to interpret what Akerlof does is to regard the explanatory power of models such as his as consisting in an ability to suggest *schemas* that allow the description of a wide variety of different and seemingly unconnected phenomena. In Akerlof's original article, for instance, the model of the second-hand car market is regarded as a mere 'finger exercise' for further application in markets as diverse as insurance, labour, other goods and credit markets. Other economists invoke transaction costs to explain the existence of firms, intergovernmental collaboration, why crime rates are higher among the poor and 'fair use' doctrines about the use of copyrighted material among many other phenomena. Many other salient theoretical concepts in economics play a similar unifying role.

Philip Kitcher developed the idea that to explain a phenomenon means to derive a description of the phenomenon from an instance of an argument pattern, instances of which can be used for deriving descriptions of many different kinds of phenomena into a formal account (Kitcher 1989). Despite its intuitive appeal, however, it is quite clear that unification cannot be all there is to explanation. How could we tell whether those factors that are salient in a model are the ones that drive the results in the real world? This is a particular problem in economics as many of the concepts that do the alleged explanatory work in models such as Akerlof's are not very discriminatory. There is hardly any market transaction that is not characterized by asymmetric information because it is virtually always the case that one party knows more or something different about a contractually relevant property. Similar observations can be made about other concepts such as human capital or transaction costs or imperfect information. In some situations such factors will be the ones that drive the result, in others they will merely provide a background against which other factors operate. But this is a dominantly qualitative question that should be decided by empirical means, not by means of models alone. Moreover, it seems unlikely that unification is necessary for explanation. Many economic events will be explained with reference to very local and idiosyncratic processes such as wars, innovations and individuals' decisions that lack the power to unify whole classes of events (for further criticism of the unification model, see Woodward 2003, ch. 8).

Nevertheless, unification plays at least two important roles in economic explanation. First, models such as Akerlof's suggest factors that may be causes of real phenomena. Unifying model schemas thus have an important heuristic role. Second, unifying explanations are in some sense desirable explanations. Even though the causal role a factor plays in bringing about a phenomenon is that which makes a model that describes the operation of this factor explanatory, it is its ability to systematize our beliefs and to reduce the number of 'brute facts' we have to accept as given that makes the explanation attractive to economists.

Equilibrium Explanation

Economics is full of equilibrium notions such as the Nash equilibrium, evolutionarily stable equilibrium, sunspot equilibrium, partial and general equilibrium theories. For a variety of reasons, economists tend to downplay explanatory accounts if these accounts do not have a bearing on theory (Heckman 2000, p. 85):

> Applications of this approach [that aims at the statistically analysis of 'natural experiments'] often run the risk of producing estimates of causal parameters that are difficult to interpret. Like the evidence produced in VAR [vector autoregression] accounting exercises, the evidence produced by this school is difficult to relate to the body of evidence about the basic behavioural elasticities of economics. The lack of a theoretical framework makes it difficult to cumulate findings across studies, or to compare the findings of one study with another. Many applications of this approach produce estimates very similar to biostatistical 'treatment effects' without any clear economic interpretation.

Equilibrium explanations, of course, have exactly this virtue: they show how some phenomenon can be systematized in a theoretical framework. Equilibrium explanations obtain at a level in between aggregate probabilistic explanations and causal-mechanical explanations. Unlike aggregate explanations, they are always formulated in terms of micro entities such as preferences, production possibilities and so forth. But, unlike causal–mechanical explanations, they rarely specify the exact details of how an equilibrium is reached or how an economy moves from one equilibrium to another. Equilibrium explanations abstract from the causal dynamics and focus on static end-points. Elliott Sober therefore argues (Sober 1983, p. 204, emphasis in original):

> Equilibrium explanations present *disjunctions* of possible causal scenarios; the actual cause is given by one of the disjuncts, but the explanation doesn't say which.

Because of this, equilibrium explanations are more unifying than explanations that describe the actual causal mechanism that lead to the equilibrium. Nevertheless, equilibrium explanations (at least in economics) tend to cite a lot of information about causes such as preferences, productivity growth, technology and so on. Although abstracting from *some* causal detail, equilibrium explanations can thus safely be regarded as a species of causal explanation.

The greatest challenge for equilibrium explanations is, however, to meet the desideratum of empirical adequacy. In order to derive any results in an equilibrium model, usually a large number of highly distorting idealizations have to be made: consumers maximize utilities and producers their profits; they operate under perfect information; markets clear instantaneously; goods are infinitely divisible and so on and so forth. Furthermore, results derived from a model making such idealizations tend to be very sensitive to specification changes. There is therefore little reason to believe that those forces that drive the equilibrium results obtain also outside the model. Hence, unless it can be shown that this is the case, equilibrium models should be regarded as mere potential explanations.

Conclusion: The Variety of Causal Explanations

The different types of explanation perform different epistemological roles. Very detailed causal–mechanistic explanations can be contrastive: they can provide information about what is special about the way in which a phenomenon came about, the way in which its causal history differs from the causal histories of other, similar phenomena. Aggregate and unifying explanations, by contrast, are comparative: they provide information about what similar or different phenomena have in common (cf. Pettit 1993, pp. 253–7). For those who are interested in explanation mostly for the practical goals of prediction and policy, aggregate explanations will often be the relevant type. However, mechanistic knowledge can be used to improve predictions, for instance, because it may provide information about the ways in which aggregate relationships sometimes fail to hold. Those with more purely cognitive goals will often prefer explanations that unify.

See Also

- ▶ Causality in Economics and Econometrics
- ▶ Models

Bibliography

Akerlof, G. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.

Cartwright, N. 1983. *How the laws of physics lie*. Oxford: Oxford University Press.

Cartwright, N. 1999. *The dappled world*. Cambridge: Cambridge University Press.

Dupré, J. 1984. Probabilistic causality emancipated. In *Midwest studies in philosophy*, vol. 9, ed. P. French, T. Uehling, Jr. and H. Wettstein. Minneapolis: University of Minnesota Press.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.

Friedman, M., and A. Schwartz. 1963. Money and business cycles. *Review of Economics and Statistics* 45: 32–64.

Heckman, J. 2000. Causal parameters and policy analysis in economics: A twentieth century perspective. *Quarterly Journal of Economics* 115: 45–97.

Hempel, C., and P. Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of Science* 15: 135–175.
Kitcher, P. 1989. Explanatory unification and the causal structure of the world. In *Scientific explanation*, ed. P. Kitcher and W. Salmon. Minneapolis: University of Minnesota Press.
Mill, J.S. 1830. *Essays on some unsettled questions of political economy*. London: Parker.
Pettit, P. 1993. *The common mind*. Oxford: Oxford University Press.
Salmon, W. 1971. Statistical explanation. In *Statistical explanation and statistical relevance*. Pittsburgh: Pittsburgh University Press.
Salmon, W. 1984. *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
Sober, E. 1983. Equilibrium explanation. *Philosophical Studies* 43: 201–210.
van Fraassen, B. 1980. *The scientific image*. Oxford: Clarendon.
Woodward, J. 2003. *Making things happen*. Oxford: Oxford University Press.

Exploitation

Gilbert L. Skillman

Abstract

This term 'exploitation' is used to characterize social relationships in which one party takes advantage of the attributes or position of another. Used normatively, it conveys the stronger sense that the exploiter takes *inappropriate* or *unfair* advantage of another's condition. While the concept has been invoked in a number of economic contexts, it has been treated most extensively in Marxist analysis of class relations in market economies, featuring in particular *orthodox, neo-Ricardian,* and *rational-choice Marxist* approaches to the phenomenon. There is ongoing debate concerning both the systemic basis and the normative significance of exploitation.

Keywords

Capitalism; Class; Engels, F.; Equality of opportunity; Exploitation; Feudalism; Fundamental Marxian theorem; Labour power vs. labour; Labour theory of value; Leontief production processes; Marx's analysis of capitalist production; Means of production; Neo-Ricardian economics; Orthodox Marxist analysis; Profit; Rational-choice Marxist analysis; Roemer, J.; Slavery; Sraffa, P.; Surplus; Surplus labour; Surplus value; Transformation problem

JEL Classifications

B41

As applied to social settings, 'exploitation' characterizes relationships in which one party takes advantage of the attributes or situation of another. A complete positive account of a given instance of exploitation would include, in addition to the identities of the parties to the relationship, three elements: the process by which one party exploits the other, the nature of advantage taken, and the conditions that make exploitation possible.

Used normatively, the term conveys the stronger connotation that the exploiter takes *inappropriate* or *unfair* advantage of another's condition. Understood in this latter sense, exploitation theory thus provides an alternative to the Pareto–utilitarian principle as a basis for assessing social outcomes, inasmuch as a given relationship may be deemed exploitative even if it improves the welfare of the exploited party.

The concept of exploitation has been invoked across a range of distinct economic contexts and paradigms. It has been applied, for example, in mainstream economic analysis of monopsonistic wage practices, and in feminist economics to characterize gender relations within households. The term has been treated most extensively, however, in Marxist political economy, which treatment is therefore the focus of the remaining discussion.

Marxian economic theorizing with respect to this phenomenon can, at some risk of oversimplification, be traced through three stages of development corresponding to distinct systems of analysis brought successively to bear on the question: *orthodox Marxist* analysis grounded in Karl Marx's reformulation of the classical labour theory of value; *Sraffian* or *neo-Ricardian* analysis

based on the mathematical properties of linear production systems; and *rational-choice Marxist* analysis utilizing formal methods of optimization and equilibrium analysis. The ensuing clash in perspectives both within the Marxian framework and between the latter and the liberal tradition undergirding the mainstream paradigm has generated a lively debate on the positive and normative significance of exploitation.

Marx's Account of Capitalist Exploitation

Karl Marx analysed societies in terms of relations between *classes*, defined in terms of ownership and control of alienable productive assets or *means of production.* In class systems such as feudalism or slavery, Marx held, those who controlled the means of production directly exploited workers by compelling them to expend more labour than that necessary to meet their own consumption needs. Capitalism, in Marx's account, is a specific historical form of class society in which antagonistic class relations are mediated by market transactions. The central analytical problem, in Marx's view, is thus to explain how exploitation of labour might arise in individually voluntary exchanges between traders with formally equal property rights.

Marx's solution to this problem, developed in the first volume of *Capital* (1867), is grounded in the postulate, previously advanced by David Ricardo, that a commodity's *value* is determined by the labour time necessary to produce it under average production conditions. On the basis of this value framework, Marx advanced three propositions: (*a*) capitalist profit is based on the extraction of surplus labour, and thus the exploitation of workers; (*b*) in industrial capitalism, the locus of exploitation is the capitalist-directed production process rather than the marketplace, although exchange relations are the necessary pretext for exploitation to occur; and (*c*) the systemic basis for exploitation under capitalism is that workers are 'free in the double sense' , that is, both legally able to offer their services in the labour market and 'free' of owning any substantial means of production themselves – a condition Marx saw as the outcome of historical processes of expropriation such as the enclosure movement in pre-industrial Great Britain.

Marx argued the first proposition, which in later formulations has been termed the 'fundamental Marxian theorem' (FMT), on the heuristic premise that commodity prices are proportional to their respective labour values. The key to Marx's demonstration is his distinction between *labour power*, that is, the capacity for productive effort, sold as a commodity in capitalist labour markets, and *labour*, the exercise of that capacity. If commodities exchange at their respective values, then positive profit is possible if and only if the labour expended by workers in capitalist production exceeds the value of their labour power – that is, workers perform surplus labour for capitalist firm owners.

Marx acknowledged that commodities typically do not exchange at their values, but maintained that prices were nonetheless 'regulated' by values, and attempted in the third volume of *Capital* (edited and published in 1894 after Marx's death by his collaborator Frederick Engels) to demonstrate the aggregate correspondence of profit and surplus labour given competitive market valuation of commodities resulting in economy-wide equalization of profit rates. His treatment of this *transformation problem* is suspect, however, due to an evidently inconsistent 'transformation' of value magnitudes in his expressions determining market prices.

Marx's second proposition emerged as a corollary of his distinction between labour power and labour. Marx argued that what capitalists purchase in the market is not a specified amount of labour time or bundle of labour services, but rather just individuals' capacity to work for a given period of time. The use value of this capacity must therefore be realized by extracting surplus labour in the process of capitalist-controlled production. It should be noted, however, that Marx did not insist as a matter of *definition* that capitalist exploitation occurs in the arena of production. In earlier drafts of *Capital* (including that which provided the basis for Engels's edition of the third volume), he frequently alluded to instances of exploitation arising from transactions involving the finance of

commodity production without direct capitalist supervision.

With respect to the third proposition, Marx argued in his theory of economic colonization that workers would be unwilling to provide capitalists with surplus value on a regular basis if they were able to produce independently to meet their own needs. Thus, capitalist profit could not in his assessment be systematic unless workers were on the whole divested of substantial ownership in the means of production, such that their livelihoods would be significantly compromised if they chose not to work for wages.

Neo-Ricardian Refinements of the FMT

The calculation of commodity labour values, necessary for the determination of exploitation status in an exchange economy, is most straightforward in the case of *Leontief* production processes, characterized by fixed input coefficients and the absence of jointly produced goods, a scenario consistent with Marx's account in the first volume of *Capital* and his treatment of the transformation problem in the third volume. However, even with this basic representation of production conditions, competitively determined commodity prices are typically disproportionate to their values. In light of his problematic 'transformation' procedure, Marx's account raises a question concerning the generality of the FMT.

Beginning in the 1960s, this question was taken up by a number of economists who, following a mode of inquiry introduced by Sraffa (1960), applied mathematical results characterizing the formal properties of linear systems understood to represent unit input requirements in a capitalist economy based on Leontief production. This literature is informed by an important theorem due independently to Frobenius and Perron (see Kurz and Salvadori 1995) which states in effect that, if it is possible to produce a surplus net of physical input requirements and real wages, and if all commodities require the direct or indirect input of other commodities, then there exists a competitive price vector supporting a strictly positive rate of profit. On the basis of this result, one can establish the FMT, framed as the formal equivalence of positive profit and the exploitation of labour given any vector of positive competitively determined commodity prices. This equivalence was subsequently extended to scenarios involving joint production and the presence of multiple alternative Leontief techniques for producing each commodity, subject to a reformulation of labour values, to be discussed below.

The Sraffian approach prompted a number of debates in the Marxian literature. An overriding concern is methodological in nature, as orthodox Marxists have questioned the relevance of ahistorical mathematical models to Marx's analytical methods and insights. There are, moreover, substantive correlates of their methodological concerns. On the one hand, the neo-Ricardian approach is silent with respect to Marx's second proposition, as the formal representation of production conditions central to the new framework simply takes as given the quantitative translation of labour capacity into productive labour expenditures. Similarly, the question of the systemic conditions enabling the extraction of surplus labour is necessarily begged in this analysis. Consequently, the equivalence between profit and capitalist exploitation established in the new versions of the FMT does not lend itself readily to an assessment of the causal connection between surplus labour and profit; the asserted equivalence might, for example, simply be a reflection of some unspecified underlying condition.

On the other hand, the Sraffian approach challenges the explanatory primacy of labour values, in so far as competitive prices are seen to be calculable directly from production and distribution conditions without any intermediate derivation of labour values. Furthermore, generalization to the case of multiple techniques and joint production introduces an element of ambiguity in the standard calculation of individual commodity values because multiple imputations of labour expenditure are then possible for each individual commodity. Morishima (Morishima and Catephores 1978) addressed the latter problem by redefining labour values in terms of the minimum possible direct labour required to produce a given bundle of net outputs. This procedure

preserves the FMT, but at a potential cost of empirical relevance, as it is at best not obvious that actual capitalists would pursue this objective in selecting among alternative techniques.

While thus casting doubt on Marx's hypothesis that individual commodity prices are somehow regulated by their respective labour values, the Sraffian approach does not directly challenge Marx's labour-based definition of exploitation. Although it provides a basis for calculating economic surplus without reference to labour values, and indicates that the operation of a viable competitive economy is consistent with a range of class distributions of that surplus, this framework establishes no independent basis for judging any such distributive outcome to be the consequence of exploitation.

Rational-choice Marxism and the Systemic Basis of Exploitation

Although the orthodox Marxian account of capitalist exploitation is clearly not methodologically individualist in nature, presumably its exponents understand it to be consistent with the interactions of self-seeking individuals responding intelligently to their available options. Granting the latter possibility raises plausible questions as to why, on the one hand, workers allow themselves to be persistently exploited by capitalists, or on the other, why capitalists choose to exercise direct control over production rather than using simple contractual means for extracting surplus labour. The corresponding normative concern asks how exploitation might be considered morally objectionable if understood to occur in voluntary transactions between rational individuals.

These and similar questions motivate a body of inquiry that has come to be termed 'rational-choice Marxism', which utilizes mainstream tools of optimization and equilibrium analysis in investigating the systemic basis and features of exploitation in market economies. A central point of reference for this stage of the Marxian literature is Roemer's *General Theory of Exploitation and Class* (1982), which poses a cogent and fundamental challenge to the canonical Marxian account. Roemer proposes a re-conceptualization of economic exploitation that dispenses entirely with labour value calculations and makes explicit the systemic basis for moral objections to its existence.

Roemer investigates the traditional labour value-based approach to exploitation in the context of both subsistence and accumulating exchange economies, initially maintaining the standard Leontief specification of production conditions and calculation of commodity values in terms of direct and indirect unit labour requirements. In this sense, his analysis commences where the Sraffian approach leaves off, embedding its representation of production conditions in a general equilibrium framework with optimizing agents.

Roemer expands this framework in two ways. First, in the scenario of subsistence exchange economies, markets for produced commodities are alternatively combined with labour and credit markets facilitating the organization of production activity. Second, in the context of accumulating market economies, he allows for unequal endowments of labour capacity and a more general representation of production possibilities.

Roemer's argument is chiefly organized around three analytical results roughly corresponding to Marx's original propositions. First, the *class-exploitation correspondence principle* (CECP) reflects Marx's fundamental assessment of the class basis of exploitation: 'capitalists', that is, those who hire or extend credit to workers, exploit, while those who work for wages or borrow to finance production activities are exploited. In the strong version of this theorem, class status is furthermore consistently linked to the market value of alienable endowments, such that the sufficiently wealthy become exploiting capitalists and the relatively indigent become exploited workers in market equilibrium.

The CECP deepens and extends the sense of the FMT by linking exploitation status to class position. Roemer demonstrates that the CECP obtains for both subsistence and accumulating economies given Leontief technology and identical preferences and labour endowments, subject only to modifications in the labour-based index of

exploitation status to accommodate the distinction between subsistence and surplus economies.

Second, Roemer argues that neither the degree nor the structure of equilibrium exploitation depend in general on whether production activities are supported by labour or credit markets. At first glance, this *isomorphism theorem* appears to rebut Marx's identification of production as the primary locus of exploitation. However, as in the Sraffian approach, Roemer's framework abstracts from the problem of translating labour capacity into labour performed. Thus the implied rebuttal of Marx's second proposition is contingent: on the assumption that capitalists face no significant obstacles in contracting for given labour services or loan repayments with interest, exchange rather than production relations provide the essential framework for capitalist exploitation. This qualified refutation nonetheless carries theoretical force, in so far as Marx's value-theoretic account of capitalist exploitation in the first volume of *Capital* does not explicitly feature conditions that would in contemporary terms be described as contracting failures.

Third, Roemer's general equilibrium framework affords a characterization of the systemic conditions for exploitation to occur in either subsistence or accumulating exchange economies. Roemer identifies *differential ownership of scarce productive assets* (DOSPA) as the necessary and (absent significant contracting failures) sufficient basis for the existence of exploitation. Differential ownership refers to inequality in the market value of individuals' endowments of productive assets, specifically (given the assumption of identical labour capacities) their endowments of *alienable* productive assets, that is, 'capital'. Significantly, the sense of *capital scarcity* in Roemer's framework is that of *absolute* scarcity, such that the market demand for capital assets (either direct or imputed) is positive at the level of total available capital supply. Otherwise, since Roemer's actors don't discount future payoffs or demand risk premia, capital goods would not command a positive rate of return.

It is readily seen why wealth inequality and capital scarcity are jointly necessary for exploitation to arise. If productive assets are equally distributed, then even if capital is scarce no agent wishes to supply it because of the resulting sacrifice in increased labour expenditure. If capital is not scarce in the absolute sense, then no agent demands it even if others enjoy greater wealth. This result presumes that individuals are identical except in their endowments of alienable assets.

Realistic generalizations of production and endowment conditions, Roemer argues, undermine the sense and scope of Marx's labour-based theory of exploitation. First, representing production possibilities as a convex cone, which maintains the assumption of constant returns to scale but allows for such practical features as factor substitution, fixed capital and joint production, introduces a tension between the labour-based conceptions of commodity value and capitalist exploitation: even if Morishima's revised valuation procedure were adopted to accommodate these more inclusive production conditions, the CECP no longer obtains in general.

To address this problem, Roemer suggests a plausible reformulation of labour value in which capitalists, rather than adopting the labour-minimizing technique for given net outputs, are instead understood to choose techniques which minimize costs at going market prices. As he demonstrates, this modification preserves the CECP, but at the necessary cost of rendering commodity values strictly dependent on commodity prices, thus reversing the causal connection asserted by Marx.

The CECP is more fundamentally compromised by quantitative and qualitative variations in individuals' inalienable endowments of labour capacity. Merely quantitative variations in labour endowments, corresponding for example to the case that some workers are able to work harder or faster than others, weakly preserve the CECP, but eliminate the monotonic relationship between class status and alienable wealth. Thus, from the standpoint of Marx's canonical account, perversities can arise such as rich but exploited wage workers or poor but exploiting capital suppliers. (A similar difficulty arises, it might be noted, given additional non-produced factors such as 'land,' which further disrupt the connection between embodied labour times and individual wealth.)

Heterogeneous labour endowments, reflecting differential abilities across distinct production tasks, raise an even more fundamental problem with respect to defining embodied labour magnitudes, as qualitatively distinct labour expenditures must somehow be aggregated before individuals' exploitation status can be assessed. But, even if a consistent procedure for aggregating heterogeneous labour inputs were identified, such as weighting them with their respective wage rates, no meaningful and consistent relationship among wealth, class position and exploitation status can thereby be established in general. Thus, the labour value-based conception of exploitation appears to break down entirely in this realistic scenario.

The Normative Significance of Exploitation

These anomalies prompt Roemer to advance a new conception of exploitation that assesses economic outcomes through comparison with some alternative property rights regime reflecting a given distributional norm. Besides sidestepping the inconsistencies arising in the labour-based approach to exploitation, this formulation offers the advantage of making explicit the grounds for judging given relationships as exploitative. As such it has potential applications beyond the scope of Marxian analysis.

Roemer's approach can be represented in terms of a cooperative game with transferable payoffs. Let N denote a set of economic actors with representative subset C, called a *coalition.* Let (C_i, C_j) designate disjoint subsets of N representing distinct coalitions in the larger society. (In Roemer's account, these coalitions are defined more specifically as complements, so that $C_j = N - C_i$.) Suppose that coalitional payoffs in this game are determined by the function $u(C)$, and define a withdrawal rule $\pi(C)$ for the game, interpreted as the payoff to coalition C should it choose to reject the allocation $u(C)$ and freely withdraw with the resources permitted it by the game's rules.

Then, given the payoff allocation rule $u(C)$, coalition C_i is said to be exploited by coalition C_j relative to withdrawal rule $\pi(C)$ if (*i*) $u(C_i) < \pi(C)$ (C_i would be made better off by withdrawing), (*ii*) $u(C_j) > \pi(Cj)$ (C_j would be made worse off by withdrawing), and (*iii*) C_j *dominates* C_i. Roemer doesn't specify the meaning of 'domination' in the *General Theory,* and offers alternative formulations in subsequent work without settling on a definitive statement of the condition. A plausible interpretation is that the payoffs $(u(C_i), u(C_j))$are somehow imposed by C_j in lieu of a feasible alternative payoff rule $\tilde{u}(C)$ such that $\tilde{u}(C_i) \geq \pi(C_i)$(coalition C_i isn't made better off by withdrawing, given $\tilde{u}$) and $u(C_j) > \tilde{u}(C_j)$ (coalition C_j benefits by choosing u over $\tilde{u}$).

An allocation $u(C)$ is then called 'non-exploitative' (NE) relative to the game characterized by $\pi(C)$ if no coalition $C \subset N$ is exploited. NE allocations are thus related to the well-known concept of the *core* of a game characterized by $\pi(C)$. All allocations in the core are NE, but not necessarily vice versa, since an allocation may be NE yet Pareto-inefficient.

The normative significance of this formulation clearly depends in part on the specification and justification of the withdrawal rule $\pi(C)$. Roemer posits three forms or levels of exploitation corresponding to different types of economic organization. *Feudal* exploitation is defined relative to a rule allowing individuals to withdraw with the free use of their existing endowments, the interpretation being that this form of exploitation involves coercive or strategic strictures on others' utilization of their property. *Capitalist* exploitation is said to exist in the context of a rule allowing actors to withdraw with the per capita share of the economy's alienable productive assets, reflecting the assessment that wealth inequality is a prerequisite for Marxian exploitation in private ownership economies. Finally, *socialist* exploitation is defined relative to a rule allowing individuals to withdraw with the per capita share of both *alienable* and *inalienable* productive assets, including skills and innate abilities.

In subsequent work (for example, Roemer 1985), Roemer has argued that moral objections to exploitation are most appropriately understood as normative demands for the provision of equal

economic opportunities to all. This is a plausible inference, and one that may in any case be compelling to those who characterize capitalist economic relations as exploitative. However, in light of the dominance condition in Roemer's three-part definition, one might with equal justification condemn the *relational* aspect of exploitation, specifically the acts by which exploiters take advantage of the vulnerable position of others, whatever the material conditions that created this asymmetry.

This point can be illustrated with reference to two strands in the rational-choice Marxist literature subsequent to Roemer's *General Theory.* One involves the criticism that Roemer's use of the competitive general equilibrium framework abstracts from contracting failures and thus ignores crucial aspects of the process by which workers are exploited in capitalist economies. In their influential treatment of this issue, Bowles and Gintis (1990) argue that the strategic response of capital suppliers to contracting failures in labour markets creates a situation of *contested exchange* in which a form of unilateral economic power is exercised over workers in capital-owned firms. They argue for the institution of democratically organized firms as a safeguard against the exercise of such power, without qualifications as to the distributional basis for this form of exchange.

A second line of criticism concerns Roemer's use of a static analytical framework in treating the essentially dynamic phenomenon of capital accumulation. This point is developed by Veneziani (2006), who argues that the condition of absolute capital scarcity, necessary for the existence of exploitation in Roemer's models, becomes extremely fragile once embedded in a genuinely intertemporal equilibrium context. On this basis, Veneziani suggests that the competitive model premised on perfect contracting arrangements is not a suitable vehicle for illuminating the Marxian theory of exploitation.

The validity of this claim, and more generally the systemic basis for capital scarcity and profit, remain open theoretical and empirical questions that stand at the boundary distinguishing Marxian and mainstream economic perspectives. In any case, there are distinct senses in which enduring capital scarcity might be deemed consistent with the manifestation of exploitative economic relationships. One possibility is that capital scarcity is somehow preserved by unequal material conditions. If, for example, time or risk preferences were income elastic, then persistent capital scarcity might arise from inequalities of the sort targeted in Roemer's account. However, even if such preferences were innate and generally uncorrelated with wealth, consistent with the standard neoclassical account, exploitation might still be said to arise if capital suppliers used inappropriate means (such as those criticized by Bowles and Gintis) to maximize the advantage derived from their unique position.

See Also

- ▶ Equality of Opportunity
- ▶ Marx's Analysis of Capitalist Production
- ▶ Neo-Ricardian Economics

Bibliography

Bowles, S., and H. Gintis. 1990. Contested exchange: New microfoundations for the political economy of capitalism. *Politics and Society* 18: 165–222.

Goodin, R. 1987. Exploiting a situation and exploiting a person. In *Modern theories of exploitation*, ed. A. Reeve. London: Sage Publications.

Kurz, H., and N. Salvadori. 1995. *Theory of production: A long-period analysis*. Cambridge: Cambridge University Press.

Marx, K. 1867. *Capital*: *A critique of political economy,* vol. 1. London: Penguin Books, 1990.

Marx, K. 1894. *Capital*: *A critique of political economy,* vol. 3, ed. F. Engels. London: Penguin Books, 1991.

Morishima, M., and G. Catephores. 1978. *Value, exploitation and growth*. Maidenhead: McGraw-Hill.

Nielsen, K., and R. Ware. 1997. *Exploitation*. Atlantic Highlands: Humanities Press.

Roemer, J. 1982. *A general theory of exploitation and class*. Cambridge: Harvard University Press.

Roemer, J. 1985. Should Marxists be interested in exploitation? *Philosophy and Public Affairs* 14: 30–65.

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

Veneziani, R. 2006. Exploitation and time. *Journal of Economic Theory* (forthcoming).

Extended Family

Olivia Harris

It has long been assumed that extended families are typical of pre-capitalist or non-capitalist societies, while the nuclear family form is the product of industrialization and urbanization. Modernization theories, deriving ultimately from nineteenth-century thinkers such as the French social reformer Frédéric Le Play (e.g., 1871), and finding different forms of expression in the Chicago School of urban sociology (e.g., Wirth 1938) and Parsonian functionalism (e.g., Parsons and Bales 1955), was articulated in a moderate form by W. Goode:

> Whenever the economic system expands through industrialisation, family patterns change. Extended kinship ties weaken, lineage patterns dissolve, and a trend toward some form of the conjugal system generally begins to appear -that is, the nuclear family becomes a more independent kinship unit (1963, p. 6).

While Goode himself recognizes that the conjugal family was prevalent in Western Europe long before the Industrial Revolution and limits himself to stating how functionally suited it is to the industrial system, it has long been assumed that the nuclear family emerged as a result of the development of capitalism (e.g., Tawney 1912). How this supposed transformation is interpreted depends on ideological positions: for those critical of the effects of capitalism the extended family evokes a world of solidarity and human values, while for the opposite tradition which finds its decisive expression in liberalism as a political doctrine, the extended family serves to maintain dependency between kin and to prevent the development of the entrepreneurial spirit.

What is meant by the extended family? The terms is ambiguous in the same way as the concept of the family itself: that is, it can refer either to a *co-resident group*, consisting of a wider group of kin than the single nuclear family, or to a network of genealogically and affinally related kin who cooperate and interact closely. However, in either case it is used especially to contrast with the dissolving of kin ties and their replacement by various types of voluntary association and contractual ties, which are said to be typical of capitalist society.

The extended family is frequently defined by the criterion of co-residence, not only because large residential groupings contrast so strikingly with the small units of Western capitalist society, but also because there is a normative assumption contained within the word 'family' that close kin *should* share their resources and if possible live together.

However, even the criterion of co-residence is not as clear-cut as might at first appear (Goody 1972). Available historical and anthropological evidence reveals that there are many different ways in which kin can share domestic space, from maintaining virtually independent budgets, to operating as a close-knit single economic unit with a strong head (usually known as a 'patriarch'). Moreover, while most authors try to maintain a distinction between the terms family and household, the terms frequently become elided or confused. The word family of course derives from the Latin *familia*, which referred to a whole complex household enterprise, including slaves. This broader definition of 'family' was only restricted to genealogical kin from the beginning of the nineteenth century (Flandrin 1976, pp. 4–10). The ambiguity arises from the contemporary ideology that co-resident domestic groups (i.e. households) *should* be based on close kin relationships, and that the intimacy, cooperation and pooling of resources found within a household are only appropriate between close kin. More specifically, Western familial ideology assumes that the co-resident kin group is the social unit within which sharing and pooling take place, while exchange takes place *between* households (Harris 1982).

Such assumptions have two problems: first, that they render us blind to the presence within households of people who are not related to the household core, that is, servants, lodgers and others. This has had damaging consequences for European family history, which only in recent years has begun to appreciate the significance of what is clearly a major pattern of European family

organization and domestic life (Macfarlane 1970; Harris 1982; Smith 1984). Secondly, such assumptions place undue emphasis on the individual household unit, at the expense of adequate consideration of movement and cooperation *between* households or their members.

Apart from methodological problems, the view that extended families are 'pre-capitalist' while the nuclear family is typical of capitalism has turned out to be inaccurate even for English history and the early development of capitalism. Macfarlane has recently summarized a large body of historical research to argue that the English peasant economy was not based on extended families from at least the thirteenth century. English rural society was mobile, with a developed market in land and labour; children were as likely to work as servants for a wage and buy land when they reached maturity, as to inherit a family farm (Macfarlane 1978). Moreover, a study of nineteenth-century Lancashire proposed that industrialization actually *increased* the number of extended (i.e. three-generation) households (Anderson 1971; see also Tilly and Scott 1978; Hareven 1982).

On a more general level, the work of Laslett and his associates has used extensive historical demographic research to argue that 'the nuclear family predominates numerically almost everywhere, even in underdeveloped parts of the world' (Laslett 1972, p. 9). Laslett's arguments were particularly directed against the orthodoxy established by Le Play that the three-generational 'stem family' *(famille souche)* was the dominant family form of the European peasantry, derived from factors such as the inalienability of the land belonging to a particular house and patriline, the buying out of siblings by a chosen successor, and extensive provisions for retirement.

Certain difficulties can be found with Laslett's influential and important arguments; in particular, Berkner (1972) has demonstrated the existence of stem families in eighteenth-century Austria by emphasizing an essential feature of households ignored by Laslett, namely that they change over time in accordance with individual life cycles and mortality rates. Thus, even in areas where the 'stem family' is the basic principle of organization, only a minority of actual households will conform to this type. This approach has been extended by Wolf (1984) to include the notion of 'family cycle' in a discussion of rural Taiwan in the twentieth century. Others have argued that Laslett's use of the 'community' as his basic unit of analysis is inappropriate, since there are significant variations between households according to class and socio-economic position. Overall there is a problem in assessing how far majority household forms in terms of *statistical frequency* reflect what each society considers to be the ideal. The discrepancy can be illustrated by contemporary industrial Britain, where research has revealed that a surprisingly large percentage of households do not conform to the ideal nuclear family type, for all that it is enshrined in legislation, welfare policies and religious belief (McIntosh 1979). The problems are obviously magnified when we turn to scanty historical data.

In recent years, detailed historical research on European families has revealed a complexity and variation that modifies Laslett's early argument but also refuses any simple correlation of family types with particular modes of production, economic stages, or even countries. Even regarding England, opinions differ as to how far and in what circumstances the nuclear family was the dominant type; taking a broader European perspective, Anderson summarizes the debate as follows:

> the European pre-industrial household was a regionally diverse one with England, northern France, North America and possibly the Low Countries . . . being unique in both their low proportion of complex households and their overall homogeneity of household patterns. By contrast, areas of much greater complexity predominated in the east and south . . . while in Northern Europe a more locally diverse pattern was found (1980, p. 29).

Various explanations have been proposed for variation in household and family forms; the influence of different systems of distributing productive property has rightly been considered of major importance. Goody (1976) offers a global theory of the formation of domestic groups in terms of land distribution which in turn he derives from agricultural technology (see also Goody et al. 1976). However, it is too general to be applicable to the understanding of local variations; a recent

exhaustive discussion of the evidence from English history from 1250 to 1800 concludes that it would be hard to maintain that the relationship between landed property and the family's development cycle was the sole or even the most important determinant of rural family forms (Smith 1984, p. 86).

In modernization theories, one of the structural explanations for the replacement of the extended family by the nuclear or conjugal family was precisely the shift away from agrarian production with land as the basic means of subsistence, to an industrial system in which the majority owned no means of production except their own labour power. However, some have argued that the organization of labour can be a major determinant of household size in peasant societies. Conversely, studies of proto-industrialization and factory production show how family ties can be strengthened and household size increased in order to maximize cooperation and the pooling of labour (Anderson 1971; Medick 1976). The same pattern can be documented outside of Europe in different contexts: some of the classic examples of large extended family households, for example, the Indian joint family, or the Japanese *dozuku*, have shown remarkable resilience in adapting to the various processes of urbanization and industrialization (Yanagisako 1979; see also Smith et al. 1984).

Large, extended family households, although not as generally found in pre-capitalist societies as was once supposed, do occur in many different contexts. However, there are major problems of definition. Is it really appropriate to include within a single category an Amazonian longhouse (*maloca*) inhabited by a variety of agnatically related families who cooperate in consumption, and the famous Balkan *zadruga*, where an older man might run a unit consisting of up to fifty people all his direct descendants, who operate as a single production unit?

Thus the whole notion of the extended family household needs substantial modification, whether one considers its historical distribution, its determinants, or its status as an individual unit.

In the broader sense of extended family as a network of kin, too, debates centre on how far such kin ties are typical of 'pre-capitalist' societies, and how far they disintegrate with the development of capitalism. Shorter (1975) presents a provocative version of the modernization thesis, arguing that the 'modern family' is private and more independent of wider kin and social ties than the 'traditional' family. But there is substantial disagreement: students of European history cite evidence to argue that neighbourhood ties have long been more significant in everyday life than kin ties (e.g., Macfarlane 1970). Goody (1983) argues that European family structures are unusual in this respect, because of policies of the early Church to proscribe marriage between close kin. Conversely, case studies from the nations of the economic periphery, and of non-European migrant groups in metropolitan regions, indicate how suited extended family networks are to business success (e.g., the classic studies for West Africa of Cohen, 1969, and Okali, 1983). Overall, the emphasis of modernization theories on linear change determined by the economy cannot be sustained, both because the pattern supposed to be typical of industrial society is found much earlier in European history, and because it is tied too closely to a supposed functional fit with industrial production. With the current restructuring of the world economy away from industry, we can expect extended family forms to thrive in many economic situations.

See Also

- ▶ Demographic Transition
- ▶ Economic Anthropology
- ▶ Family
- ▶ Inheritance

Bibliography

Anderson, M. 1971. *Family structure in nineteenth century Lancashire*. Cambridge: Cambridge University Press.

Anderson, M. 1980. *Approaches to the history of the Western family 1500–1914*. London: Macmillan.

Berkner, L. 1972. The stem family and developmental cycle of the peasant household: An eighteenth-century Austrian example. *American Historical Review* 77: 398–418.

Flandrin, J-L. 1976. *Families in former times*. Trans. Cambridge: Cambridge University Press, 1979.
Goode, W. 1963. *World revolution and family patterns*. New York/Glencoe: Free Press.
Goody, J. 1972. The evolution of the family. In *Household and family in past time*, ed. P. Laslett and R. Wall. Cambridge: Cambridge University Press.
Goody, J. 1976. *Production and reproduction: A comparative study of the domestic domain*. Cambridge: Cambridge, University Press.
Goody, J. 1983. *The development of the family and marriage in Europe*. Cambridge: Cambridge University Press.
Goody, J., J. Thirsk, and E.P. Thompson (eds.). 1976. *Family and inheritance*. Cambridge: Cambridge University Press.
Hareven, T. 1982. *Family time and industrial time*. Cambridge: Cambridge University Press.
Harris, O. 1982. Households and their boundaries. *History Workshop Journal* 13: 143–152.
Laslett, P. 1972. Introduction. In *Household and family in past time*, ed. P. Laslett and R. Wall. Cambridge: Cambridge University Press.
Le Play, F. 1871. L'organisation de la famille selon le vrai modèle signalé par l'histoire de toutes les races et de tous les temps. Paris.
Macfarlane, A. 1970. *The family life of Ralph Josselin*. Cambridge: Cambridge University Press.
Macfarlane, A. 1978. *The origins of English individualism*. Oxford: Basil Blackwell.
McIntosh, M. 1979. The welfare state and the needs of the dependent family. In *Fit work for women*, ed. S. Burman. London: Croom Helm.
Medick, H. 1976. The proto-industrial family economy. *Social History* 1(3): 291–315.
Netting, R., R. Wilk, and E. Arnould (eds.). 1984. *Households. comparative and historical studies of the domestic group*. Berkeley: University of California Press.
Parsons, T., and R. Bales. 1955. *Family. Socialization and interaction process*. Glencoe: Free Press.
Shorter, E. 1975. *The making of the modern family*. New York: Basic Books.
Smith, R. (ed.). 1984. *Land, kinship and life cycle*. Cambridge: Cambridge University Press.
Smith, J., I. Wallerstein, and H.D. Evers (eds.). 1984. *Households and the world economy*. California: Sage Publications.
Tawney, R.H. 1912. *The agrarian problem of the sixteenth century*. London: Longmans.
Tilly, L., and J. Scott. 1978. *Women, work and family*. New York: Holt, Rinehart & Winston.
Wirth, L. 1938. Urbanism as a way of life. *American Journal of Sociology* 44(July): 1–24.
Wolf, A. 1984. Family life and the life cycle in rural China. In *Households. Comparative and historical studies of the domestic group*, ed. R. Netting, R. Wilk, and E. Arnould. Berkeley: University of California Press.
Yanagisako, S. 1979. Family and household: The analysis of domestic groups. *Annual Review of Anthropology* 8: 161–205.

Extensive and Intensive Rent

Guido Montani

The distinction between extensive and intensive rent appears clearly in the history of economic thought with Ricardo, even though a number of economists discussed these concepts previously on various occasions (e.g. Anderson 1777). After Ricardo, until the end of the century, every economist understood the concept of rent to mean the possibility of obtaining an income from the ownership of scarce natural resources, such as land and mines. But that notion of rent changed progressively and substantially after the so-called 'marginalist revolution'. It may be, therefore, useful to examine the transformation of the notion of rent from classical to marginalist economics.

To understand the concept of rent in classical political economy proper, it is essential to relate it to the notion of surplus, that is the quantity of commodities which at the end of the production cycle (usually the year) are left for consumption, net investment (or waste) after the means of production are replaced so that the new production cycle can begin again on at least the same scale. Quesnay was the first to show clearly that rent is a component of surplus. In the *Tableau économique* (1758) rent appears as a share of the agricultural product – agriculture is the only productive sector – which is paid every year by farmers to landlords. In the same way, Adam Smith spoke of rent as a revenue belonging to landlords as a surplus. As soon as land becomes private property [says Smith] the landlord demands a share of almost all the produce which the labourer can either raise, or collect from it. His rent makes the first deduction from the produce of the labour which is employed upon land (Smith 1776, p. 58). But neither Quesnay, nor Smith gave a satisfactory explanation of the causes affecting the level of the rates of rent. For Quesnay, landlords had a feudal right to ask farmers for a rent. And Smith considered rent as a 'monopoly price'.

Only with Ricardo, who published in 1815, at the same time as Malthus's and West's works on the same subject, *An Essay on the Influence of a Low Price of Corn on the Profits of Stock*, does the classical theory of rent take on a clear and precise shape. In a certain sense, the *Principles*, published two years later, can be considered, as far as rent is concerned, only as an application of the labour theory of value of the specific case already worked on in the *Essay*. 'Rent', says Ricardo, 'is that portion of the produce of the earth, which is paid to the landlord for the use of the original and indestructible powers of the soil' (Ricardo 1817, p. 67). The landlord can raise a rent only when land becomes scarce, that is when the demand for agricultural produce cannot be satisfied without putting into cultivation lands of inferior quality.

> When in the progress of society [says Ricardo] land of the second degree of fertility is taken into cultivation, rent immediately commences on that of the first quality, and the amount of that rent will depend on the difference in the quality of these two portions of land. (Ricardo 1817, p. 70)

This is extensive rent.

But Ricardo recognized a second kind of rent. Before inferior land is cultivated, 'capital can be employed more productively on those lands which are already in cultivation'. This is intensive rent, 'for rent is always the difference between the produce obtained by the employment of two equal quantities of capital and labour' (Ricardo 1817, p. 71).

This viewpoint changed considerably in the marginalist theory of value and distribution. The distinction between extensive and intensive rent was maintained, but progressively the notion of rent as a surplus disappeared and a new explanation of the law of decreasing returns was introduced into economic theory. In Jevons, for instance, rent is again a surplus, even if its explanation is founded on the new mathematical marginalist techniques. 'The accepted theory of rent', says Jevons, referring to Ricardo's and J.S. Mill's doctrine, 'needs little or no alteration to adapt it to expression in mathematical symbols'. And supposing that a worker is employed on a given area of land, rent will be 'the excess of produce which can be exacted from him… if he be not himself the owner of the land' (Jevons 1871, pp. 220, 223). A few years later, Walras protested over such a treatment of rent. Lesson 39 of his *Eléments d'économie politique pure* (1874) is devoted to an 'Exposition and refutation of the English theory of rent'. His starting point is Ricardo's and Mill's distinction between extensive and intensive rent on which he is in agreement, but after a mathematical restatement of these theories, he asks himself 'why this school does not try to formulate a unified general theory to determine the prices of all productive services in the same way'. Walras' aim is to include rent theory in a system of general economic equilibrium in which all prices of commodities and services are interdependent.

> All that remains of Ricardo's theory after a rigorous analysis is that rent is not a component part, but a result, of the price of products. But the same thing can be said of wages and interest. Hence, rent, wages, interest, the prices of products, and the coefficients of production are all unknowns within the same problem; they must always be determined together and not independently of one another. (Walras 1874, p. 416, 418)

Marshall considerably extended the concept of rent. 'The rent of land', affirms Marshall, 'is no unique fact, but simply the chief species of a large genus of economic phenomena' (Marshall 1890, p. 523). All income can include an element of rent. 'There is an element of true rent in the composite product that is commonly called wages, an element of true earnings in what is commonly called rent and so on' (Marshall 1890, p. 350). If the supply of a certain factor is scarce, and cannot increase in a certain period of time, then it is possible to gain an income which may be properly called rent. We can imagine cases of 'pure rent' and cases of 'quasi-rent'. An example of quasi-rent is that of incomes gained on old investments of capital. There is no sharp line of division between pure rent and quasi-rent. Commodities which are in short supply in the short run can be produced in a greater quantity in the long run, so that any possibility of obtaining rent disappears.

Moreover, Marshall softened the distinction between extensive and intensive rent, which he calls differential and scarcity rent. 'In a sense',

says Marshall, 'all rents are scarcity rents, and all rents are differential rents'. We can say that differential rent arises because the land of a single quality comes to be, at a certain point in time, in short supply.

> In this connection [concludes Marshall], it may be noted that the opinion that the existence of inferior land, or other agents of production, tends to raise the rents of the better agents is not merely untrue. It is the reverse of the truth. For, if the bad land were to be flooded and rendered incapable of producing anything at all, the cultivation of other land would need to be more intensive; and therefore the price of the product would be higher, and rents generally would be higher, than if that land had been a poor contributor to the total stock of produce. (Marshall 1890, pp. 351–2)

If Marshall was the economist who gave the greatest contribution to the extension of the concept of rent, Wicksteed was the one who gave it a precise and probably definitive arrangement inside the marginalist theory of distribution. According the Wicksteed the marginalist, or as he calls it, the differential theory of distribution, when fully grasped 'must destroy the very conception of separate laws of distribution such as the law of rent, the law of interest, or the law of wages' (Wicksteed 1914, p. 789). The possibility of coordinating the distribution shares, in order that their sum amounts to the total net product, rests on the fact that the differential service to production of every factor is always the same, even if the way may differ from factor to factor. For example, for land the quality which is relevant is extension, for labour skill and dexterity, etc. 'The law of distribution', affirms Wicksteed, 'is one, and is governed not by the differences of nature in the factors, but by the identity of their differential effect' (Wicksteed 1914, p. 789). In other words, what is important for the marginalist theory is not the heterogeneity of factors, but the differential effect of a different quantity of a factor of the same quality.

The consequences of this observation are quite radical.

> Ricardo's celebrated law of rent really asserts nothing except that the superior article fetches the superior price, in proportion to its superiority; and it is obvious that all 'superiorities' in land, whether arising from 'inalienable' properties or from expenditure of capital, tell in exactly the same way upon the rent. (Wicksteed 1914, p. 790)

When we consider the usual diagram in which different qualities of land are represented on the abscissas and the different fertilities obtained form a given 'dose' of labour and capital on the ordinate we must be aware of the fact that that curve is not a *functional curve*. We simply arranged the different kinds of land in a descending order according to their fertility. On the contrary, if we increase the quantity of a certain factor in relation to a fixed quantity of another factor we may construct a curve which shows a *functional* relation between the 'doses' of the factor and the marginal product, which has a behaviour depending on the quantity of the variable factor employed. In the first case considered, there is no law of rent at all 'but the tacit assumption that the differential theory of distribution is true of every factor of production except land, and that rent is what is left after everything that is not rent is taken away'. Only with a functional curve do we have a true theory of rent and distribution, but in such a case 'we must understand that when the differential distribution is affected there is no surplus or residuum at all' (Wicksteed 1914, pp. 791–2). All the product is therefore distributed to the production factors according to their marginal product.

To conclude our short description of the place of rent in the marginalist theory of distribution, it is clear that with Wicksteed rent is no longer a share of the annual surplus, but an income, the nature of which is perfectly symmetrical with that of capital and labour: they are all paid according to their marginal product. But to reach that result it was essential to reduce the classical law of decreasing returns to the 'law of variable proportions'. As Wicksteed states firmly, only the expansion or contraction of a homogeneous factor in relation to a given quantity of another one gives rise to a marginal product.

The marginalist theory of rent and, more generally, of distribution were widely accepted until recent times. Only with the publication in 1960 of *Production of Commodities by Means of Commodities* by Piero Sraffa were increasing doubts

raised by economists on the reasonableness of its assumptions. Sraffa resumed the classical point of view of the problem of value and distribution, placing the notion of surplus at the centre of his inquiry. Inside that theoretical framework it is again quite natural to consider rent as a surplus, that is, a share of the net income distributed to landlords (or other owners of scarce natural resources). Indeed, Sraffa states that 'it is hardly necessary to dwell on the doctrine that 'taxes on rent fall wholly on landlords and thus cannot affect the prices of commodities or the rate of profits' (Sraffa 1960 p. 74), which was precisely the point of view defended by Ricardo.

Sraffa draws a clear-cut distinction between extensive and intensive rent. If we consider a system of production equations, we can take into consideration two sectors: industry and agriculture. The equations of the industrial sector enable us to determine the rate of profits, given the wage rate, and all industrial prices. In the agricultural sector only corn is produced (which is not utilized as a means of production in the industrial sector). Let us suppose now that n different qualities of land are disposable. If the quantity of corn required can be raised on the more fertile land, land will be redundant and there will be no rent. The price of corn is determined by its production equation, since the costs of its means of production (reckoned on the basis of the industrial prices) are known and the rate of profits and wages are equally known. Only when the need arises to put less fertile lands into cultivation will a rent be possible for the owners of the more fertile farms. But the marginal land will pay no rent.

Intensive rent is possible when land of a single quality is in short supply. In such a case, two different processes or methods of cultivation can be used side by side determining a uniform rent per acre. The existence of two methods adopted simultaneously on the same land should be considered as the result of a process of intensive diminishing returns.

> The existence side by side, of two methods can be regarded as a phase in the course of a progressive increase of production on the land. The increase takes place through the gradual extension of the method that produces more corn at a higher unit cost, at the expense of the method that produces less. (Sraffa 1960, p. 76)

It is worthwhile here noticing that Sraffa's analysis is based on the assumption that the quantities of commodities which should be produced and brought to the market are given. Any inquiry regarding the factors affecting these magnitudes should be undertaken at a further stage: here relative prices, the levels of personal income, consumption habits, and so on should be considered as independent variables affecting the quantities of commodities required. But if it is agreed, as a first step, to consider the quantities to be produced as given, some interesting results will emerge:

(a) In the case of extensive rent, the extension of production from a more fertile to a less fertile land, owing to the need to produce more agricultural products, will cause a lowering of the general rate of profits (if corn is a means of production), an increase in the price of the agricultural commodity in relation the industrial ones and a consequent rise in rents on the more fertile lands. In a similar way, in the case of intensive rent, an increase in the quantity produced of a certain agricultural good will cause a change in the methods of production – the old pair of techniques will give place to a new more efficient one – with a consequent increase in the agricultural price and rent, and the reduction of the general rate of profits. Both results (Montani 1972) are perfectly in agreement with the Ricardian doctrine of rent.

(b) The order of fertility of the various kinds of land is not given, once and for all, by nature. The more fertile lands (i.e. lands which are put into cultivation first because of the rate of profits they give to the agricultural entrepreneur) do not coincide with the lands paying higher rents. Ricardo's opinion is not correct on this point:

> When land of the third quality is taken into cultivation, rent immediately commences on the second. At the same time, the rent of the first quality will rise, for that must always be above the rent of the second, by the difference between the produce

> which they yield with a given quantity of capital and labour (Ricardo, p. 70; our italics).

Generally speaking, if lands 1, 2 and 3 are already cultivated and rent on land 1 is higher than rent on land 2, it may happen that when land 4 is put into cultivation the rate of rent on land 2 may become greater than the rate of rent on land 1. The reversal of the order of rents is possible both for the rate of rent per unit of land and for rate of rent per unit of product (Montani 1972).

> (a) The scarcity of land does not depend only on the quantity of the agricultural product to be produced, but on the distribution of income between profits and wages too. Given certain methods of production, some natural resources, as land, mineral deposits, etc, become 'scarce' or 'redundant' according to the quantity produced of a given commodity and to the relative level of the rate of profits in relation to wages. This may happen both for extensive and intensive rent. Therefore, since no change in the proportions of the 'production factors' occurs when the natural resource becomes scarce or redundant owing to the change in distribution, it is obvious that the meaning of 'decreasing returns' inside the classical theory of rent (and distribution) is different from that connected with the 'law of variable proportions.' (Montani 1975)

The vicissitudes of the theory of rent in the history of economic thought are, therefore, strictly connected to the meaning of the law of decreasing returns; this can be appreciated from what has been said above about how the content of this law changed considerably during the transition from the classical to the marginalist paradigm. Marshall was well aware of the diversity of the two points of view, and stated quite clearly that

> the diminishing return which arises from an ill-proportioned application of the various agents of production into a particular task has little in common with that broad tendency to the pressure of a crowded and growing population on the means of subsistence. The great classical Law of Diminishing Return has its chief application, not to any one particular crop, but to all the chief food crops. (Marshall, p. 338)

This classical meaning of the law was progressively forgotten by the economists owing to the over-narrow view imposed by the marginalist theory of distribution.

See Also

- ▶ Absolute Rent
- ▶ Land Rent
- ▶ Rent
- ▶ Ricardo, David (1772–1823)

Bibliography

Anderson, J. 1777. An Inquiry into the Nature of the Corn-Laws; with a View to the New Corn-Bill proposed for Scotland. Edinburgh.

Jevons, W.S. 1871. *The theory of political economy*. Harmondsworth: Penguin, 1970.

Marshall, A. 1890. *Principles of economics*, 8th ed. London: Macmillan, 1972.

Montani, G. 1972. La teoria ricardiana della rendita. *L'Industria* 3/4: 221–243.

Montani, G. 1975. Scarce natural resources and income distribution. *Metroeconomica* 27: 68–101.

Ricardo, D. 1817. On the principles of political economy and taxation. In *The works and correspondence of David Ricardo*, vol. I, ed. P. Sraffa. Cambridge: Cambridge University Press, 1960. Reprinted, 1966.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: Everyman's Library, 1964.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

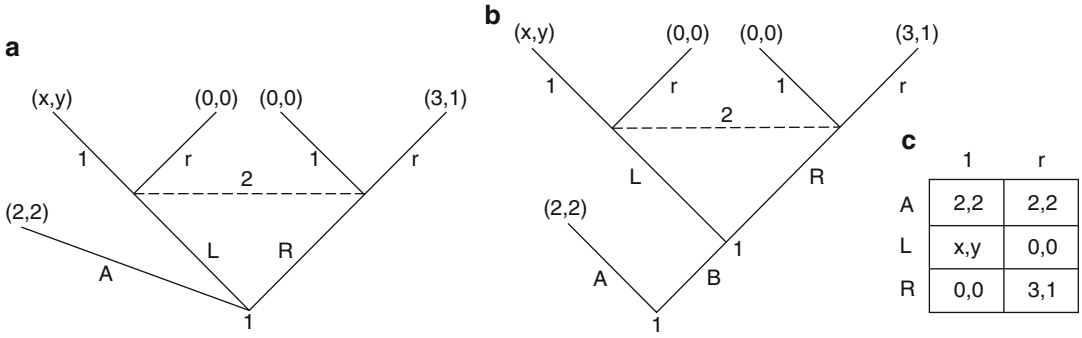
Walras, L. 1874. *Eléments d'économie politique pure*. Trans. as *Elements of pure economics*, ed. W. Jaffé, 1954. Reprinted, Fairfield: A.M. Kelley, 1977.

Wicksteed P.H. 1914. The scope and method of political economy in the light of the 'marginal' theory of value and distribution. *Economic Journal* 24, March. Reprinted in the *The common sense of political economy and selected papers and reviews on economic theory*, Vol. II, ed. L. Robbins, 1933. Reprinted, New York: A.M. Kelley, 1967.

Extensive Form Games

Eric Van Damme

The most general model used to describe conflict situations is the extensive form model, which specifies in detail the dynamic evolution of each situation and thus provides an exact description of 'who knows what when' and 'what is the



Extensive Form Games, Fig. 1

consequence of which'. The model should contain all relevant aspects of the situation; in particular, any possibility of (pre)commitment should be explicitly included. This implies that the game should be analysed by solution concepts from noncooperative game theory, that is, refinements of Nash equilibria. The term extensive form game was coined in von Neumann and Morgenstern (1944) in which a set theoretic approach was used. We will describe the graph theoretical representation proposed in Kuhn (1953) that has become the standard model. For convenience, attention will be restricted to finite games.

The basic element in the Kuhn representation of an n-person extensive form game is a rooted tree, that is, a directed acyclic graph with a distinguished vertex. The game starts at the root of the tree. The tree's terminal nodes correspond to the endpoints of the game and associated with each of these there is an n-vector of real numbers specifying the payoff to each player (in von Neumann–Morgenstern utilities) that results from that play. The nonterminal nodes represent the decision points in the game. Each such point is labelled with an index $i (i \in \{0, 1, \ldots, n\})$ indicating which player has to move at that point.

Player O is the chance player who performs the moves of nature. A maximal set of decision points that a player cannot distinguish between is called an information set. A choice at an information set associates a unique successor to every decision point in this set, hence, a choice consists of a set of edges, exactly one edge emanating from each point in the set. Information sets of the chance player are singletons and the probability of each choice of chance is specified. Formally then, an extensive form game is a sixtuple $\Gamma = (K, P, U, C, p, h)$ which respectively specify the underlying tree, the player labelling, the information sets, the choices, the probabilities of chance choices and the payoffs.

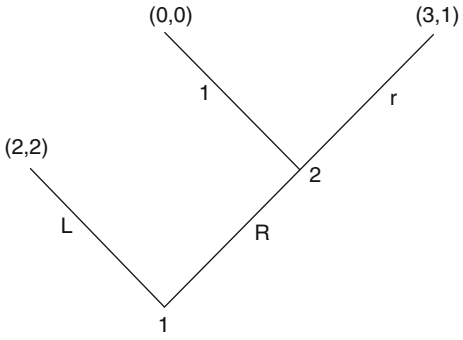
As an example, consider the 2-person game of Fig. 1a. First player 1 has to move. If he chooses A, the game terminates with both players receiving 2. If he chooses L or R, player 2 has to move and, when he is called to move, this player does not know whether L or R has been chosen. Hence, the 2 decision points of player 2 constitute an information set and this is indicated by a dashed line connecting the points. If the choices L and ℓ are taken, then player 1 receives x, while player 2 gets y. The payoff vectors at the other endpoints are listed similarly, that is, with player l's payoff first. The game of Fig. 1b differs from that in Fig. 1a only in the fact that now player 1 has to choose between L and R only after he has decided not to choose A. In this case, the game admits a proper subgame starting at the second decision point of player 1. This subgame can also be interpreted as the players making their choices simultaneously.

A strategy is a complete specification of how a player intends to play in every contingency that might arise. It can be planned in advance and can be given to an agent (or a computing machine) who can then play on behalf of the player. A pure strategy specifies a single choice at each information set, a behaviour strategy prescribes local randomization among choices at information sets and a mixed strategy requires a player to randomize

several pure strategies at the beginning of the game. The normal form of an extensive game is a table listing all pure strategy combinations and the payoff vectors resulting from them. Figure 1c displays the normal form of Fig. 1a, and, up to inessential details, this also represents the game of Fig. 1b. The normal form suppresses the dynamic structure of the extensive game and condenses all decision-making into one stage. This normalization offers a major conceptual simplification, at the expense of computational complexity: the set of strategies may be so large that normalization is not practical. Below we return to the question of whether essential information is destroyed when a game is normalized.

A game is said to be of perfect recall if each player always remembers what he has previously known or done, that is, if information is increasing over time. A game may fail to have perfect recall when a player is a team such as in bridge and in this case behaviour strategies may be inferior to mixed strategies since the latter allow for complete correlation between different agents of the team. However, by modelling different agents as different players with the same payoff function one can restore perfect recall, hence, in the literature attention is usually restricted to this class of games. In Kuhn (1953) and Aumann (1964) is has been shown that, if there is perfect recall, the restriction to behaviour strategies is justified.

A game is said to be of perfect information if all information sets are singletons, that is, if there are no simultaneous moves and if each player always is perfectly informed about anything that happened in the past. In this case, there is no need to randomize and the game can be solved by working backwards from the end (as already observed in Zermelo 1913). For generic games, this procedure yields a unique solution which is also the solution obtained by iterative elimination of dominated strategies in the normal form. The assumption of the model that there are no external commitment possibilities implies that only this dynamic programming solution is viable; however, this generally is not the unique Nash equilibrium. In the game of Fig. 2, the roll-back procedure yields (R, r), but a second equilibrium is (L, ℓ). The latter is a Nash equilibrium since



Extensive Form Games, Fig. 2

player 2 does not have to execute the threat when it is believed. However, the threat is not credible: player 2 has to move only when 1 has chosen R and facing the *fait accompli* that R has been chosen, player 2 is better off choosing r. Note that it is essential that 2 cannot commit himself: If he could we would have a different game of which the outcome could perfectly well be (2, 2).

A major part of noncooperative game theory is concerned with how to extend the backwards induction principle to games with imperfect information, that is how to exclude intuitively unreasonable Nash equilibria in general. This research originates with Selten (1965) in which the concept of subgame perfect equilibria was introduced, that is, of strategies that constitute an equilibrium in every subgame. If $y < 0$, then the unique equilibrium of the subgame in Fig. 1b is (R, r) and, consequently, (BR, r) is the unique subgame perfect equilibrium in that case. If $x < 2$, however, then (AL, ℓ) is an equilibrium that is not subgame perfect. The game of Fig. 1a does not admit any proper subgames; hence, any equilibrium is subgame perfect, in particular (A, ℓ) is subgame perfect if $x > 2$. This shows that the set of subgame perfect equilibria depends on the details of the tree and that the criterion of subgame perfection does not eliminate all intuitively unreasonable equilibria.

To remedy the latter drawback, the concept of (trembling hand) perfectness was introduced in Selten (1975). The idea behind this concept is that with a small probability players make mistakes, so that each choice is taken with an infinitesimal probability and, hence, each information

set can be reached. If $y \leq 0$, then the unique perfect equilibrium outcome in Fig. 1a, b is (1, 3): player 2 is forced to choose r since L and R occur with positive probability.

The perfectness concept is closely related to the sequential equilibrium concept proposed in Kreps and Wilson (1982). The latter is based on the idea of 'Bayesian' players who construct subjective beliefs about where they are in the tree when an information set is reached unexpectedly and who maximize expected payoffs associated with such beliefs. The requirements that beliefs be shared by players and that they be consistent with the strategies being played (Bayesian updating) imply that the difference from perfection is only marginal. In Fig. 1a, only (R, r) is sequential when $y < 0$. When $y = 0$, then (A, ℓ) is sequential, but not perfect: choosing ℓ is justified if one assigns probability 1 to the mistake L, but according to perfectness R also occurs with a positive probability.

Unfortunately, the great freedom that one has in constructing beliefs implies that many intuitively unreasonable equilibria are sequential. In Fig. 1a, if $y > 0$, then player 2 can justify playing ℓ by assigning probability 1 to the 'mistake' L, hence, (A, ℓ) is a sequential equilibrium if $x \leq 2$. However, if $x < 0$, then L is dominated by both A and R and thinking that 1 has chosen L is certainly nonsensical. (Note that, if $x < 0$, then (AL, ℓ) is not a sequential equilibrium of the game of Fig. 1b, hence, the set of sequential (perfect) equilibrium outcomes depends on the details of the tree.) By assuming that a player will make more costly mistakes with a much smaller probability than less costly ones (as in Myerson's (1978) concept of proper equilibria) one can eliminate the equilibrium (A, ℓ) when $x \leq 0$ (since then L is dominated by R), but this does not work if $x > 0$. Still, the equilibrium (A, ℓ) is nonsensical if $x < 2$: If player 2 is reached, he should conclude that player 1 has passed off a payoff of 2 and, hence, that he aims for the payoff of 3 and has chosen R. Consequently, player 2 should respond by r: only the equilibrium (R, r) is viable.

What distinguishes the equilibrium (R, r) in Fig. 1 is that this is the only one that is stable against all small perturbations of the equilibrium strategies, and the above discussion suggests that such equilibria might be the proper objects to study. An investigation of these stable equilibria has been performed in Kohlberg and Mertens (1984) and they have shown that whether an equilibrium outcome is stable or not can already be detected in the normal form. This brings us back to the question of whether an extensive form is adequately represented by its normal form, that is, whether two extensive games with the same normal form are equivalent. One answer is that this depends on the solution concept employed: it is affirmative for Nash equilibria, for proper equilibria (van Damme 1983, 1984) and for stable equilibria, i.e. for the strongest and the weakest concepts, but it is negative for the intermediate concepts of (subgame) perfect and sequential equilibria. A more satisfactory answer is provided by a theorem of Thompson (1952) (see Kohlberg and Mertens 1984) that completely characterizes the class of transformations that can be applied to an extensive form game without changing its (reduced) normal form: The normal form is an adequate representation if and only if these transformations are inessential. Nevertheless, the normal form should be used with care, especially in games with incomplete information (cf. Harsanyi 1967–8; Aumann and Maschler 1972), or when communication is possible (cf. Myerson 1986).

See Also

- ▶ Cooperative games
- ▶ Game theory
- ▶ Games with incomplete information
- ▶ Nash equilibrium
- ▶ Non-cooperative games

Bibliography

Aumann, R.J. 1964. Mixed and behavior strategies in infinite extensive games. In *Advances in game theory*, ed. M. Dresher, L.S. Shapley, and A.W. Tucker. Princeton: Princeton University Press.

Aumann, R.J., and M. Maschler. 1972. Some thoughts on the minimax principle. *Management Science* 18(5): 54–63.

Harsanyi, J.C. 1967–8. Games with incomplete information played by 'Bayesian' players. *Management*

Science 14; Pt I, (3), November 1967, 159–182; Pt II, (5), January 1968, 320–334; Pt III, (7), March 1968, 486–502.
Kohlberg, E., and J.-F. Mertens. 1984. On the strategic stability of equilibria. Mimeo, Harvard Graduate School of Business Administration. Reprinted in *Econometrica* 53, 1985, 1375–1385.
Kreps, D.M., and R. Wilson. 1982. Sequential equilibria. *Econometrica* 50(4): 863–894.
Kuhn, H.W. 1953. Extensive games and the problem of information. *Annals of Mathematics Studies* 28: 193–216.
Myerson, R.B. 1978. Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 7(2): 73–80.
Myerson, R.B. 1986. Multistage games with communication. *Econometrica* 54(2): 323–358.
Selten, R. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift für die gesamte Staatswissenschaft* 121: 301–324; 667–689.
Selten, R. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4(1): 25–55.
Thompson, F.B. 1952. *Equivalence of games in extensive form*, RAND Publication RM–769. Santa Monica: Rand Corp.
Van Damme, E.E.C. 1983. *Refinements of the Nash equilibrium concept*, Lecture Notes in Economics and Mathematical Systems No. 219. Berlin: Springer-Verlag.
Van Damme, E.E.C. 1984. A relation between perfect equilibria in extensive form games and proper equilibria in normal form games. *International Journal of Game Theory* 13(1): 1–13.
Von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Zermelo, E. 1913. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. *Proceedings of the Fifth International Congress of Mathematicians* 2: 501–504.

External Debt

Peter M. Oppenheimer

The term ‘external debt’ refers to financial obligations incurred by individuals or, more commonly, institutions resident in one country vis-à-vis those resident in another. In other words, the obligations cross the borders of sovereign states. Usually different nationalities or citizenships are involved, as well as different residencies, but this is not strictly necessary. For instance, a US corporation in the United States may borrow from a US bank branch on London, which in turn funds the loan by taking deposits from American residents of the United States. The involvement of a financial intermediary in London means that this chain of transactions among American nationals is governed in part by English law and regulations rather than, or as well as, US law (and of course the laws may conflict). In fact, legal or regulatory differences are likely to be the reason why such all-American transactions move ‘off-shore’ in the first place. For example, reserve or liquidity requirements applying to US banks in London may be less severe than those applying in New York, thus enabling the London branch to offer more attractive terms than its New York counterpart to depositors and borrowers alike.

In the above example US residents have both a debt and a claim vis-à-vis non-residents. So the net external asset position of the United States is the same as it would be if neither of these particular transactions had taken place. The same would apply if the bank in London were British or Japanese rather than American, and if US residents had placed their overseas assets not with this bank but with some other institution in a third country. Nonetheless, with countries as with individuals or firms, it is not only the net (external) financial position which matters. A country (or a firm) may have substantial net (external) assets and still face financial difficulties – a liquidity crisis or a run on the currency – if its liabilities are more liquid than its assets, and external creditors demand repayment; or if domestic residents wish to send more capital abroad and foreign creditors refuse to grant additional loans. Conversely, insofar as a debtor country is readily able to meet the interest due on its foreign debt from export proceeds or other normal receipts, its net debtor position is not a source of embarrassment.

These considerations point the way to theoretical and practical criteria for external borrowing and lending by a country; but a number of other points have first to be clarified. Aside from debts, external liabilities (and assets) may take the form of claims on real property (ordinary shares, land, buildings, etc). Unlike such claims, debts carry

legal obligations to pay interest and/or amortization over some stated period; but, unless specifically indexed, these obligations are in nominal money terms, not in real terms. The rate of interest may be either fixed or fluctuating according to some formula, but will normally depend in large measure upon the currency denomination of the loan. The faster the real-terms purchasing power of a currency is expected to depreciate in the future, the higher the interest rate that it will carry. Creditors therefore lose, and debtors gain, to the extent that the currency in question loses real value during the period of the loan at a faster pace than financial markets had anticipated at the time the loan was made; and conversely when it loses real value more slowly. Monetary authorities in countries with longer-term borrowings denominated in their own currency are accordingly exposed to a degree of temptation (moral hazard) to encourage faster depreciation of their currencies in order to reduce the burden of such debts; and conversely for monetary authorities of countries with longer-term external lending in their own currency (though this has not been viewed as a significant risk in practice).

External debt may take the form either of securities that are in principle marketable (bonds or bills) or of non-marketable claims such as bank loans, trade credit between firms or government-to-government loans. Before 1914, international lending was conducted overwhelmingly by way of securities; subsequently other claims have played a larger role, most especially so in the 1970s, which saw a huge wave of lending to less developed countries (notably in Latin America) in the form of medium-term credits, denominated mainly in US dollars, put together by syndicates of major commercial banks.

External Borrowing and the Macroeconomy

Starting with a clean sheet, a country acquires *net* external liabilities (assets) by spending abroad out of current income more (less) than it is currently earning abroad – in other words, by running a deficit (surplus) on the current account of its balance of payments. Once external assets and liabilities exist, the net position is also affected by valuation changes, i.e. appreciation and depreciation, of the various items. Another way of looking at a current payments deficit is to say that it represents external dis-saving by the country in question, and comes about as a result of the combined saving and spending decisions of all the country's residents. Macroeconomic policies exercise a crucial influence, both directly (the government being a major spending unit) and indirectly (through the impact of fiscal, monetary and exchange-rate policies upon private-sector outlays). It is important to distinguish between external saving and total saving. Most of a country's total saving is matched by expenditure on domestic capital formation, i.e. investment in buildings, machinery, etc., whether by the private or public sectors. External dis-saving and hence the current payments deficit equals the *excess* of domestic investment over national saving (and conversely for external saving and a current payments surplus). In symbols $X - M = S - I$, where X is export receipts, including earnings on overseas assets as well as sales of goods and services, M is import payments, S is national saving and I is domestic investment. Clearly the net decline in the external asset/liability position which matches a current account deficit may take many different forms – such as acquisition of domestic factories by foreign firms, sale of overseas assets by domestic residents, drawing down of official foreign-exchange reserves, etc. – and need not involve external borrowing in the strict sense of the term.

Criteria

Some external borrowing is undertaken in emergency or disequilibrium situations, as an alternative to depletion of foreign exchange reserves or, beyond that, to rapid elimination of a current account deficit by means of restraint on domestic expenditure which would cause significant disruption or hardship. Thus, a domestic investment boom may be sustained and allowed to run its course rather than be cut short by balance-of-payments pressures. Or the impact of an

unwelcome event, such as a sudden and unexpected drop in export receipts in a primary-product-exporting country, may be cushioned. In this instance the argument for emergency borrowing is most evident when the drop in export income is expected for good reason to be short-lived. When it is expected to be of long duration, some external borrowing may still be appropriate, so that the country may adapt to its new circumstances gradually. In sum, borrowing is rational whenever the discounted present value of its cost is judged to be less than that of the expenditures which would be foregone if the borrowing did not take place.

Analogous criteria apply to external borrowing along an equilibrium growth path. Assume for simplicity that there is a single world-wide interest rate at which funds can be both lent and borrowed. A country should undertake all domestic investment projects expected to be profitable in foreign-currency terms at this interest rate, arranging the structure and timing of its investment programme so as to maximize the present-value difference between total return and total cost. At the same time the country's total saving will depend on population size, income level, income distribution and thriftiness or time preference. The country will be a net borrower or net lender abroad in any particular phase of its development, depending on whether domestic investment is running ahead of or behind national saving. More explicit and precise formulations of this principle can be derived in the context of the mathematical theory of optimum growth (see Shell 1967).

Uncertainty, Default Risk and the Banking System

The principles outlined so far largely abstract from problems of imperfect foresight and information. Any credit market has to cope with such problems, but special issues arise in relation to international lending.

Although a borrower promises to service his debt, circumstances or dishonesty may lead him to default on or repudiate the obligation. In the face of this possibility, various kinds of safeguard are available to actual or potential creditors. First, private creditors should not merely lend to one or a few borrowers, but should spread risks by holding a market portfolio of claims. Secondly, the market itself will generate differential pricing of loans, with debtors judged riskier having to pay higher interest rates and/or being rationed in the amounts that they can borrow at any price. The role of rationing is emphasised by the theory of adverse selection (Stiglitz and Weiss 1981), which suggests that higher interest rates tend to drive away the prudent and good-quality borrowers, leaving mainly the more doubtful prospects still in the market. Thirdly, borrowers who default on their obligations are liable to forfeit property either automatically or after legal proceedings. Such proceedings may, however, be costly and troublesome to pursue, especially in the international context, where action may be required in more than one country. Fourthly, default or bankruptcy carries a stigma which will prevent or hamper the possessor's access to credit and other markets for a greater or lesser period of time. By the same token, creditors may be able to enforce prudent policies of their choosing upon debtors as a condition of prolonging credit rather than declaring the debtor in default. In practice such power can be exercised only by banks or other financial institutions; and in the case of government ('sovereign') borrowers mostly only by other governments or supranational institutions such as the International Monetary Fund (IMF).

The point about continued access to international markets has to carry a lot of weight in the case of sovereign borrowers, where the mechanism of foreclosure (property forfeiture) is inapplicable. With a commercial (non-sovereign) loan, lenders will wish to ensure that the market value of the assets available as security in the event of default or insolvency is at least equal to the discounted present value of the lost interest and repayments. The nearest comparable condition in the case of a sovereign loan is to ensure that the borrower believes his self-interest to lie in maintaining debt service, because default would bring him more losses than gains. If his losses are defined simply in terms of denial of

future loan inflows (a plausible albeit somewhat narrow definition), then the incentive to default arises if and only if the discounted present value of the potential future inflows is less than that of the interest and amortization payments avoided by defaulting. To prevent this condition being met, lenders must put appropriate limits on the growth of their loans to sovereign borrowers, the limits depending at any one time on the existing level of debt, the real rate of interest and the debtor economy's growth rate (see, for instance, Niehans 1986; Eaton and Gersovitz 1981).

Suppose, however, that lenders miscalculate and allow external debt to build up to a point where sovereign borrowers are seriously tempted to default. To lure them away from this option, lenders must then continue to re-lend debt service payments as they fall due and, in addition, ensure that the real interest rate being charged is at least no higher, and probably somewhat lower, than the borrower's long-term economic growth rate. Depending on circumstances, lenders may prefer to cut their losses by accepting default; and some may have no effective choice in the matter – for instance, widely scattered small bondholders with no practical means of influencing the scale or terms of new lending.

The foregoing optimal lending strategies to sovereign borrowers may not be achievable by competitive private lenders acting on their own. Financial markets, being characterized by incompleteness of information about the creditworthiness of borrowers and about future economic trends, are prone to be heavily influenced by herd instinct and fashion, even if decisions are rationalized in analytical terms. When circumstances look favourable, it is difficultto prevent excessive growth of lending to favoured sectors or customers, including sovereign borrowers – especially with financial institutions also competing for positions of market leadership and size. Then, when the climate turns sour, private lenders left to themselves may precipitate a financial crash by recalling loans when debtors are in no position to repay. Observing these tendencies at work in 19th-century British credit markets, Walter Bagehot (1873; see also Hirsch 1977) argued that stabilization of the financial system requires a degree of restraint on competition through some mixture of oligopolistic market structure and central-bank guidance or approval, the latter as a *quid pro quo* for protecting banks against runs by means of last-resort lending facilities.

A similar lesson emerged from global lending developments in the 1970s and 1980s. When less developed countries in Latin America and elsewhere became major international borrowers after 1970, and especially after the quadrupling of oil prices in 1973–4, most of the sovereign lending was handled by the commercial bank network of the major industrial countries, which was far more flexible than official credit channels centred on the IMF in intermediating to new patterns of international capital flows. In the process, however, lending banks, especially in the United States, made themselves vulnerable by lending amounts greatly in excess of their capital and reserves to a handful of sovereign borrowers. From 1979 onwards monetary restraint imposed by the US Federal Reserve System triggered a shift in the world financial climate whose extent and duration could scarcely have been foreseen. In particular, international real interest rates rose from a negative figure to around 7 per cent, and the terms of trade moved heavily against primary producers. After two-to-three years the burden of debt service had become unsustainable in relation to borrowers' export receipts. During 1982–3 some 35 countries were obliged to seek rescheduling of their external debts. The fact that this happened in an orderly manner, and without involving a serious chain of bank failures, was due to the action of the principal OECD-country central banks, led by the Federal Reserve, in conjunction with the IMF and the Bank for International Settlements. They buttressed commercial bank lending with official credits and, more important, exercised moral suasion to induce the banks to renew sovereign loans as they fell due, rather than seek a large-scale withdrawal of funds which would not have been feasible. This cooperative 'crisis management' was the international

equivalent of last-resort lending in a domestic banking system.

See Also

- ▶ Dependency
- ▶ International Finance
- ▶ International Indebtedness

Bibliography

Begehot, W. 1873. *Lombard street*. London: H.S. King & Co.

Eaton, J., and M. Gersovitz. 1981. *Poor country borrowing in financial markets and the repudiation issue*. Princeton Studies in International Finance No. 47. Princeton: Princeton University Press.

Hirsch, F. 1977. The Bagehot problem. *Manchester School of Economics and Political Science* 45(3): 241–257.

Niehans, J. 1986. In *Strategic planning in international banking*, ed. P. Savona and G. Sutija. London.

Shell, K. (ed.). 1967. *Essays on the theory of optimum economic growth*. Cambridge, MA: MIT Press.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71(3): 393–410.

External Economies

Peter Bohm

Abstract

External economies and diseconomies nowadays generally mean unpaid side effects of one producer's output or inputs on other producers. External economies in this sense imply as a rule that market prices in a competitive market economy will not reflect marginal social costs of production, giving rise to a 'market failure'. But, with technological external economies and diseconomies now most often replaced by the well-defined concept of externality, and with pecuniary external economies and diseconomies being synonymous with general market interdependence, external economies no longer have much of a role to play in economic analysis.

Keywords

Balanced growth; Division of labour; External economies; Externalities; Investment criteria; Knight, F. H.; Linkages; Market failure; Marshall, A.; Pecuniary external economies; Pigou, A. C.; Robertson, D. H.; Scitovsky, T.; Technological external economies; Young, A. A.

JEL Classifications

B1

'The concept of external economies is one of the most elusive in economic literature.' This is how Tibor Scitovsky began his article 'Two Concepts of External Economies' (1954). His statement is still true, and it may be added that there are at least two such concepts.

The meaning of external economies and its counterpart, external diseconomies, has changed over time. Nowadays, it is essentially synonymous with externality or external effects in the sphere of production. That is, external economies (diseconomies) or positive (negative) external effects in production are unpaid side effects of one producer's output or inputs on other producers. (As an illustrative example, we can take the case where a dam constructed by a hydroelectric power plant eliminates flooding of farmers' crop fields – external economy – or reduces the catches of fishermen downstream – external diseconomies; a producer's pollution which increases the costs of, *inter alia,* other producers is perhaps the most important case of externalities.) Sometimes, external economies also refer to unpaid side effects of or on consumption activities, but this meaning is disregarded here.

External economies in this modern sense imply as a rule that market prices in a competitive market economy will not reflect marginal social costs of production. Hence, a 'market failure' arises,

meaning that the market economy cannot attain a state of efficiency on its own. Specifically, in an otherwise 'perfect' market economy, a producer who has external economies (positive external effects) on other producers would not extend his externality generating activity, say, his output, to the point where marginal cost of production equals marginal social benefits of production, which amounts to the market value of his marginal output plus the market value of the side effect on the output of other producers.

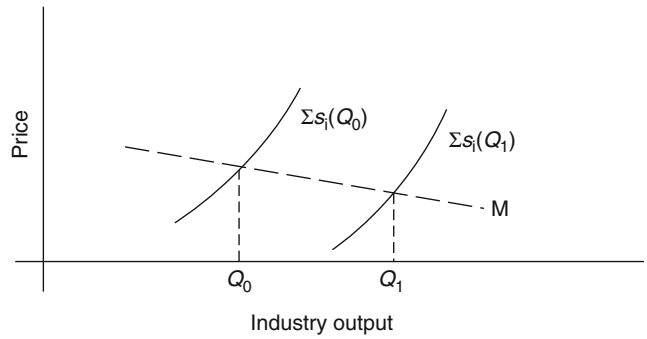
At an earlier stage, external economies in the meaning now given were called *technological external economies,* reflecting the fact that the effects were transmitted outside the market mechanism and altered the technological relationship between the recipient firm's output and the inputs under its control. Formally, we have that the output q_i, of the ith producer is affected not only by changes in his control variables, a vector x_i, but also by e_j, a variable controlled by some other producer j. This gives us the following production function:

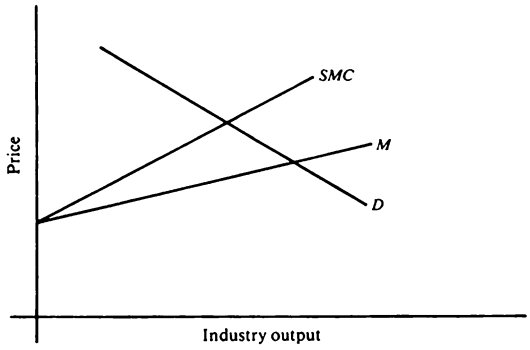
$$q_i = f_i(x_i; e_j).$$

The reason for specifying such effects as technological is that the concept of 'external economies' has been given a broad meaning ever since its introduction. During the early part of the 20th century, external economies (diseconomies) were defined so as to include beneficial (detrimental) *price effects* of producer activities. Thus, in principle, the concept included cases where increases in factor inputs by one firm lowered or raised input prices for other firms. However, much of the discussion centred on the case where increases in *industry* output lowered or raised input prices for the individual member firm. The case of reduced input prices presupposes that the supply side of the market for inputs is characterized either by imperfect competition (say, a profit-maximizing monopoly producing at decreasing marginal costs, decreasing at a rate sufficient for an increase in demand to lower price) or by a competitive industry having a downward-sloping 'supply' curve, which in turn reflects external economies in this industry. Supply conditions in the original industry, as well as in the industry producing inputs for the first industry, are shown in Fig. 1. Here, $\sum s_i(Q_0)$ is the aggregate supply of the firms in the industry when actual industry output is Q_0. When industry output increases to Q_1, input prices drop (or technological external economies arise), causing a downward shift in individual cost and supply curves and hence in the aggregate supply ($\sum s_i(Q_1)$). The curve *M,* the downward-sloping 'supply' curve, is actually a market equilibrium curve showing the equilibrium price/output combinations at different levels of demand (see Bohm 1967).

The price effects between firms or between industry and its firms were termed *pecuniary external economies* (*diseconomies*) by Viner (1931). Before him, A.C. Pigou (1920) had argued (although he phrased it differently) that

External Economies, Fig. 1





External Economies, Fig. 2

external economies and diseconomies, both technological and pecuniary, would call for government intervention in order for the industry to attain a socially efficient level of output. Specifically, Pigou argued that, if expansion of a competitive industry would increase prices of inputs sold to the industry, thus creating pecuniary external diseconomies for the individual firms in the industry, the aggregate 'supply' (or market equilibrium) curve would not reflect social marginal costs.

Pigou's argument can be illustrated in Fig. 2, where *M* is the upward-sloping long-run 'supply' curve for the industry due to rising input prices as a consequence of increasing industry output. *SMC* is the curve showing the total marginal outlay on inputs, where the difference to *M* is the increased outlay for *infra-marginal* inputs. Pigou contended (but later rescinded this position) that the *SMC* curve indicated the true social marginal costs and hence, that the price/output combination attained by the market, as shown by the intersection of market 'supply' *M* and market demand *D,* was suboptimal. He argued that the optimal level could not be attained unless a tax were levied on this industry so that, in equilibrium, price and output would be those shown by the intersection of *SMC* and *D*. (Similarly, a bounty would be required in the case of pecuniary external economies (see Fig. 1), where *SMC* would be downward-sloping and steeper than the *M* curve.)

It was demonstrated by F.H. Knight (1924) and D.H. Robertson (1924) and further elaborated by Ellis and Fellner (1943) that the total marginal outlay is irrelevant as an indicator of marginal social costs. The effect on outlay for intramarginal units of inputs represents an increment to rent on these units and hence, it is not part of the real costs of increased output; the real marginal costs are only those which have to be paid for the required marginal inputs. In other words, the rising costs for marginal inputs are those shown by the industry 'supply' curve *M*. Hence, pecuniary external economies or diseconomies would not call for government intervention. The case is different, however, for technological economies and diseconomies. Assume, for example, that output of a geographically concentrated industry pollutes the area where it is located and hence reduces the productivity of labour inputs of the individual firms in the industry. Here, all effects on input productivity, for marginal as well as intramarginal units, constitute real costs. So if the rising *M* and *SMC* reflect such effects only, Pigou's argument holds.

When the debate had arrived at this point, pecuniary external economies could be dropped as a cause of market failure and, hence, the concept lost its specific economic interest. But, now, what do the *technological* external economies and diseconomies of industry output on the individual firms in the industry actually represent? Alfred Marshall introduced the term external economies when analysing industry production costs as a function of output:

> We may divide the economies arising from an increase in the scale of production of any kind of goods, into two classes – firstly, those dependent on the general development of the industry; and secondly, those dependent on the resources of individual houses of business engaged in it, on their organization and the efficiency of their management. We may call the former *external economies,* and the latter *internal economies.* (Marshall 1920, p. 266)

The latter concept is now recognized as economies of scale in the individual firm. Marshall elaborated the meaning of the former concept using scattered examples, the most explicit of which is perhaps the increased knowledge accompanying the expansion of industry output

materialized in the publication of trade journals and other forms of improved information about markets and technology in the industry. But, in addition, he argued that industry growth, especially when concentrated to a particular region, might create a market for skilled labour, advance subsidiary industries or give rise to specialized service industries as well as improve railway communication and other infrastructure.

Thus, external economies emerged here essentially as cost reductions for individual firms as a consequence of industry growth, that is, as economies external to the firm but internal to the industry. To remain firmly within the framework of static analysis, these economies should be thought of as being reversible. But some of Marshall's examples alluded to irreversible phenomena and to dynamic effects of industry growth. This was particularly obvious when he at times referred to external economies as being dependent on the 'general progress of industrial environment'.

Marshall's claim that external economies were important in the long run – in fact, more important than internal economies – seemed to have little immediate impact on the thinking of his fellow economists. For example, the discussion of 'empty economic boxes' in the early 1920s (Clapham 1922; Pigou 1922; Robertson 1924) centred on questioning the relevance of external economies and diseconomies, the latter concept deriving from issues raised by Pigou (1920). Actually, the significance of technological external economies and diseconomies in the *static* analysis of an industry or among producers in general escaped most economists all the way up to the post-war years when traffic congestion (although already mentioned by Pigou) and the common pool problem of interdependent producers in an oil field or a fishing area became the leading examples of the industry case and environmental pollution the leading example of the general case. Eventually, diseconomies emerged as the important case and economies as the exceptional case, whereas earlier hardly any importance was attached to technological diseconomies (see, for example, Robertson 1924).

If static external effects turned out to be the most important legacy of Marshall's original contribution, external economies as a *dynamic* concept, which seems closer to Marshall's own ideas, also came to play a role in economic analysis. Dynamic external economies refer to increased division of labour resulting from industry growth, the emergence of firms specializing in new activities, some of which aim at developing capital equipment for, or servicing, other firms. An early elaboration of these ideas was made by Young (1928). Later, the concept of external economies came to play a prominent role in development planning, primarily for underdeveloped countries or regions. The general idea here was hardly at all related to the non-market interdependence of technological external economies, but rather to the market interdependence of which pecuniary external economies were part, now on an economy-wide basis. It was argued, in particular by Rosenstein-Rodan (1943) and – with a somewhat different focus – by Scitovsky (1954), that development must be planned so that supply and demand relationships among different sectors of the economy are taken into account. Specifically, it was pointed out that major investments in industrial capacity in a poor economy risk being a failure unless a required increase in the supply of inputs to meet the need of the expanding industry, as well as a sufficient stimulus of demand for the output of the expanding industry, occur at the same time. This doctrine of 'balanced growth' called for a simultaneous expansion of investment in several sectors of the economy, a 'Big Push' (Rosenstein-Rodan 1943) determined by input–output relations or general market interdependence. Scitovsky argued specifically that indivisibilities in capital formation may call for investment criteria, which – in contrast to those of individual firms – take into account the indirect supply and demand effects on profitability elsewhere in the economy.

Here, external economies came to be synonymous with what was later called linkage effects (Hirschman 1958), where backward linkages refer to the supply to the investing sector and forward linkages to the demand for the output of the investing sector. Hirschman's own analysis led

him to recommend a strategy for economic development in poor countries that was opposite to that of a 'balanced growth'. Focusing on the shortage of entrepreneurial capacity and the small impact of subtle market signals in backward areas, he advocated a strategy of 'unbalanced growth', according to which investments should be undertaken so as to reinforce market signals and create pressure for investments in sectors related to the original investments via strong linkage effects.

That linkage effects offered a more precise terminology than dynamic external economies is one reason why the latter concept now has lost ground. Another is, of course, that this meaning of external economies referred only or primarily to market interdependence, which is a general economic phenomenon. Thus, with technological external economies and diseconomies now most often replaced by the well-defined concept of external effects, and with pecuniary external economies and diseconomies being synonymous with general market interdependence, external economies no longer have much of a role to play in economic analysis. Aside from occasional use as a synonym for external effects, the concept now stands for interdependence that does not clearly fall into any of the categories mentioned here. That is, when firms affect one another in a way not covered by static equilibrium analysis or by interdependence among existing markets in the context of the dynamic analysis of economic development, external economies are still used by economists as a convenient catchall.

Again, one may ask – as did economists in the interwar period – what do these external economies actually stand for? At least some examples can be given. Growth of an industry may create a supply of new skills which turn out to provide a starting point for an altogether new line of business. Or, growth of a technologically advanced industry in a particular region leading to the location of a school of higher learning to this region may in turn stimulate – or reduce – growth of other activities. These cases are awkward to handle in traditional, well-structured economic analysis. So the main characteristic of these external economies, very much like most of those suggested by Marshall, is that we cannot yet say in any systematic way exactly what they represent.

See Also

- ▶ Externalities
- ▶ Linkages
- ▶ Young, Allyn Abbott (1876–1929)

Bibliography

American Economic Association (AEA). 1952. *Readings in price theory*, vol. 6. Homewood: Irwin.

Arndt, H.W. 1955. External economies in economic growth. *The Economic Record* 31: 192–214.

Arrow, K., and T. Scitovsky (eds.). 1969. *Readings in welfare economics.* Homewood: Irwin.

Bohm, P. 1967. *External economies in production*. Stockholm: Almqvist & Wiksell.

Clapham, J.H. 1922. On empty economic boxes. *Economic Journal* 32: 305–314. Reprinted in AEA (1952).

Ellis, H., and W. Fellner. 1943. External economies and diseconomies. *American Economic Review* 33: 493–511. Reprinted in AEA (1952).

Heller, W.P., and D.A. Starrett. 1976. On the nature of externalities. In *Theory and measurement of economic externalities*, ed. S. Lin. New York: Academic.

Hirschman, A.O. 1958. *The strategy of economic development*. New Haven: Yale University Press.

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38: 582–606. Reprinted in Arrow and Scitovsky (1969).

Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.

Meade, J. 1952. External economies and diseconomies in a competitive situation. *Economic Journal* 62: 54–67. Reprinted in Arrow and Scitovsky (1969).

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Pigou, A.C. 1922. Empty economic boxes. *Economic Journal* 32: 458–465. Reprinted in AEA (1952).

Robertson, D.H. 1924. Those empty boxes. *Economic Journal* 34: 16–30. Reprinted in AEA (1952).

Rosenstein-Rodan, P.N. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53: 202–211.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62: 70–82. Reprinted in Arrow and Scitovsky (1969).

Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3: 23–46. Reprinted in AEA (1952).

Young, A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542. Reprinted in Arrow and Scitovsky (1969).

Externalities

J. J. Laffont

Abstract

Externalities are *indirect* effects of consumption or production activity, that is, effects on agents other than the originator of such activity which do not work through the price system. In a private competitive economy, equilibria will not be in general Pareto optimal since they will reflect only *private* (direct) effects and not *social* (direct plus indirect) effects of economic activity. This article explains how this outcome arises and considers the policy responses that have been advanced to remedy the market failures stemming from externalities.

Keywords

Asymmetric information; Coalitions; Competitive equilibrium; Cooperative game theory; Environmental economics; Externalities; Imperfect information; Lump-sum transfers; Non-convexity; Pecuniary externalities; Pollution rights; Strategic behaviour; Taxation of externalities; Technological externalities

JEL Classifications

D62

Competitive equilibria are Pareto optimal when they exist if preferences are locally non-satiated and if externalities are not present in the economy. Why externalities upset the first fundamental theorem of welfare economics and which economic policies can remedy this failure are the major questions addressed below.

Technological Externalities

Let us call technological externality the indirect effect of a consumption activity or a production activity on the consumption set of a consumer, the utility function of a consumer or the production function of a producer. By 'indirect' we mean that the effect concerns an agent other than the one exerting this economic activity and that this effect does not work through the price system.

Externalities may be positive or negative and are quite diverse. Major examples include pollution activities (air pollution, water pollution, noise pollution . . .), malevolence and benevolence, positive interaction of production activities. From a practical point of view the most significant are negative pollution activities, so that we can say that the theory of technological externalities is essentially the foundation of environmental economics.

The formalization of technological externalities is achieved in microeconomics by making production sets, utility functions and production sets (or functions) affected by externalities functionally dependent on the activities of the other agents creating these indirect effects.

For example, the utility function of a consumer is made dependent on the level of production of a firm polluting the air breathed by the consumer. This modelling option that we will implicitly adopt here is right as long as the link between production and air pollution is not alterable.

If de-polluting activities are possible the link between the level of pollution and the economic activities generating them must be made explicit. An important difficulty in analysing these activities is due to the non-convexities which they usually introduce.

Pecuniary Externalities

During the 1930s, a confused debate occurred between economists on the relevance of pecuniary externalities, that is, on externalities which work through the price system. A quite general consensus was that pecuniary externalities are irrelevant for welfare economics: the fact that by increasing my consumption of whisky I affect your welfare through the consequent increase in price does not jeopardize the Pareto optimality of competitive equilibria.

This is true when all the assumptions required for the competitive equilibria to be Pareto optimal are satisfied. In such a framework prices only equate supply and demand and pecuniary externalities do not matter. As soon as we move away from this set of assumptions prices generally play additional roles. For example, in economies with incomplete contingent markets, prices span the subspace in which consumption plans can be chosen. In economies with asymmetric information, prices transmit information. When agents affect prices, they affect the welfare of the other agents by altering their feasible consumption sets or their information structures. Pecuniary externalities matter for welfare economics.

In what follows we focus only on technological externalities.

Competitive Equilibrium with Externalities

How is the characterization of Pareto optima in convex economies affected by externalities? Very simply, as Pigou early understood. The classical equality of marginal rates of substitution and marginal rates of transformation must now be expressed using *social* marginal rates and not only *private* marginal rates as in an economy without externalities. Social marginal rates must be computed taking into account direct *and* indirect effects of economic activities. For example, the marginal cost of a polluting activity must include not only the direct marginal cost of production, but also the marginal cost imposed on the environment.

Note that Pareto optima do not exclude polluting activities, but set them at levels such that their social marginal benefit equates their social marginal cost well computed.

It is now easy to understand that in a private competitive economy, equilibria will not be in general Pareto optimal since the private decentralized optimizations of economic agents lead them to the equalization of *private and not social* marginal rates through the price system.

Markets for Pollution Rights

Consider for concreteness a firm polluting a consumer. One potential solution is to create a market for this externality. Before producing, the firm must buy from the consumer the right to pollute. If both actors were behaving competitively with respect to the price of this right, the competitive equilibrium in the economy with an extended price system would be Pareto optimal, since there is no externality left.

A number of difficulties exist with this approach. In general we cannot expect agents to behave competitively unless we are in the special case of impersonal externalities. Then, there is a fundamental non-convexity in the case of negative externalities since as a negative externality increases the production set shrinks, but there is a limit to this effect which is the zero production level. Competitive equilibria cannot then exist unless bounds are set on supplies of pollution rights. (For a positive price, a firm would like to offer an infinite amount of pollution rights and close down.)

In the above set-up, the implicit status quo was the absence of externalities. The initial rights are a clean environment: 'Polluters must pay.' We can instead give to the polluting firm the right to pollute and then ask the consumer to buy from the firm a decrease of his pollution. This different allocation of initial rights does not upset the Pareto optimality of the competitive equilibrium, but of course has distributional effects.

Taxation of Externalities

The likely strategic behaviour by agents on markets of pollution rights makes taxation of externalities the most common policy tool. The polluter must then pay for each unit of a polluting activity a tax which equals the marginal cost imposed by this activity on the other agents. The polluter then internalizes the externality and Pareto optimality is restored. If the externality is positive he must be similarly subsidized.

Note that nothing is said about the amount of taxes so obtained by the government. There is no

presumption that it is given to the polluters. In fact, the implicit assumption is that it is redistributed through lump-sum transfers which do not affect agents' behaviours (in the sense of their first-order conditions). From the point of view of Pareto optimality, the important goal is to modify polluters' behaviours.

If lump-sum transfers are not available, the budget of the government must be balanced and then goods different from the polluting activities must be taxed or subsidized to solve the ensuing second-best problem.

The major difficulty with this solution is informational.

Imperfect Information

The traditional theory of externalities has proceeded as if the regulators had complete knowledge of the economy and were therefore able to compute optimal taxes, or as if agents were not behaving strategically with respect to their private information. Very often this is not the case and the problem is to elicit this private information and use it to compute taxes, a more difficult problem.

Intuitively, the solution of what is now a second-best problem is to have taxes which depend nonlinearly on polluting activities. This nonlinearity may sometimes take the extreme form of a zero tax up to a given amount and a very large tax above, a mechanism which is equivalent to a quota.

Planning and Externalities

Externalities are not only a problem of market economies with an insufficient number of markets. One way to suppress an externality between two agents is to have them integrate into a single agent. All externalities would be internalized if the whole economy was integrated.

If we leave aside imperfect information and the associated strategic behaviours, the planning problem of these integrated agents is more complicated than if externalities were not present. Planning procedures appropriated to externalities have been provided.

Externalities and Cooperative Game Theory

Suppose we attempt to represent the outcome of cooperation in an economy with externalities by the core. The core is the set of allocations which are not blocked by any coalition. A coalition blocks an allocation if it can do better for all its members than this allocation.

Externalities introduce a difficulty in the definition of a blocking coalition.

When a group of agents envision forming a coalition they must conjecture what will be the behaviour of the complementary coalition since it is affected by the externalities of this complementary coalition.

Two extreme notions have been proposed. In the a-core a coalition is said to block an allocation if it can do better, whatever the actions of the complementary coalition. This is extremely prudent. In the β-core a coalition is said to block an allocation if, for any action of the complementary coalition, it can do better. The β-core is of course included in the α-core.

Results depends a lot on these conjectures about the actions of the complementary coalition, an unsatisfactory feature. One lesson, however, is that the core may be empty, that is, that externalities introduce an element of instability in economic games.

Historical Note

Following the pioneering work by Sidgwick (1887) and Marshall (1890), Pigou (1920) has provided the basic theory of static technological externalities. Coase (1960) has explained how initial rights could be assigned in various ways. Arrow (1969) has explained how externalities could be internalized by the creation of additional markets. Starrett (1972) has pointed out the associated problem of nonconvexity. The first theorem

of existence of an equilibrium with externalities has been provided by McKenzie (1955). Shapley and Shubik (1969) studied the core with externalities. A large number of authors have studied various second-best problems associated with externalities (Buchanan 1969; Plott 1966; Diamond 1973; Sandmo 1975).

See Also

- ▶ Clubs
- ▶ Coase Theorem
- ▶ External Economies
- ▶ Incentive Compatibility

Bibliography

Arrow, K. 1969. The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In *The analysis and evaluation of public expenditures: The PPB system*, ed. Joint Economic Committee. Washington, DC: Government Printing Office.

Buchanan, J.M. 1969. External diseconomies, corrective taxes and market structure. *American Economic Review* 59: 174–176.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Diamond, P. 1973. Consumption of externalities and imperfect corrective pricing. *Bell Journal of Economics and Management Science* 4: 526–538.

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

McKenzie, L. 1955. Competitive equilibrium with dependent consumer preferences. In *Proceedings of the second symposium in linear programming*, ed. H.A. Antosiewicz. Washington, DC: National Bureau of Standards.

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Plott, C.R. 1966. Externalities and corrective taxes. *Economica* 33: 84–87.

Sandmo, A. 1975. Optimal taxation in the presence of externalities. *Swedish Journal of Economics* 77: 96–98.

Shapley, L., and M. Shubik. 1969. On the core of an economic system with externalities. *American Economic Review* 59: 687–689.

Sidgwick, H. 1887. *Principles of political economy*, 2nd ed. London: Macmillan.

Starrett, D. 1972. Fundamental non-convexities in the theory of externalities. *Journal of Economic Theory* 4: 180–199.

Extremal Quantiles and Value-at-Risk

Victor Chernozhukov and Songzi Du

Abstract

This article examines the theory and empirics of extremal quantiles in economics, in particular value-at-risk. The theory of extremes has gone through remarkable developments and produced valuable empirical findings since the late 1980s. We emphasize conditional extremal quantile models and methods, which have applications in many areas of economic analysis. Examples of applications include the analysis of factors of high risk in finance and risk management, the analysis of socio-economic factors that contribute to extremely low infant birthweights, efficiency analysis in industrial organization, the analysis of reservation rules in economic decisions, and inference in structural auction models.

Keywords

Bootstrap; Extremal conditional quantiles; Extremal quantiles; Gamma variables; Inference; Maximum likelihood; Production frontiers; Quantile regression function; Value-at-risk

JEL Classifications

D62

Introduction

Some Basics

Let a real random variable Y have a continuous distribution function $F_Y(y) = \text{Prob}[Y \leq y]$. A τ-quantile of Y is the number $F_Y^{-1}(\tau)$ such that $\text{Prob}\left[Y \leq F_Y^{-1}(\tau)\right] = \tau$ for some $\tau \in (0,1)$ (More generally, let $F_Y^{-1}(\tau) = \inf\{y : F_Y(y) > \tau\}$). The quantile function $F_Y^{-1}(\tau)$, viewed as a function of

probability index τ, is the inverse of the distribution function $F_Y(y)$. The quantile function is therefore a complete description of the distribution.

Let X be a vector of regressor variables. Let $F_Y(y|x) = \text{Prob}[Y \leq y|x]$ denote the conditional distribution function of Y given $X = x$. The conditional τ-quantile of a random variable Y with a continuous conditional distribution function is the number $F_Y^{-1}(\tau|x)$ such that $\text{Prob}\left[Y \leq F_Y^{-1}(\tau|x)|X = x\right] = \tau$. The conditional quantile function $F_Y^{-1}(\tau|x)$ viewed as a function of x is called the τ-quantile regression function. The main use of the quantile regression function $F_Y^{-1}(\tau|x)$ is to measure the effect of covariates on outcomes, both in the centre and in the upper and lower tails of an outcome distribution. To this effect, a quantile or a conditional τ-quantile will be referred to as *extremal* whenever the probability index τ is either low, $\tau \leq 0.15$, or high, $\tau \geq 0.85$. Without loss of generality, we focus the discussion on the low quantiles.

Examples as a Motivation

There are many applications of extremal quantiles in economics, particularly of extremal conditional quantiles. Here we give a sample of these applications as a motivation for what follows.

Example 1 Conditional Value-at-Risk Value-at-risk analysis seeks to forecast or explain very low conditional quantiles $F_Y^{-1}(\tau|X)$ of an institution's portfolio return, Y, tomorrow, using today's available information, X (Chernozhukov and Umantsev 2001). Typically, extremal quantiles $F_Y^{-1}(\tau|X)$ with $\tau = 0.01$ and $\tau = 0.05$ are of interest. Value-at-risk analysis is a daily activity for banking and other financial institutions, as required by the Securities and Exchange Commission and the Basle Committee on Banking Supervision. As a risk measure, value-at-risk is motivated by the safety-first decision principle formalized by Roy (1952), in which one makes optimal decisions subject to the constraint that the probability of the risk of a large loss is kept small. This and similar measures are commonly used in real-life financial management, insurance, and actuarial science (Embrechts et al. 1997).

Example 2 Determinants of Very Low Birthweights In the analysis of infant birthweights, we may be interested in how smoking, absence of prenatal care, and other types of maternal behaviour affect various birthweights (Abrevaya 2001). Of special interest, however, are the very low quantiles, since low birthweights have been linked to subsequent health problems. Chernozhukov (2006) provides an empirical study of extreme birthweights.

Example 3 Probabilistic Production Frontiers An important form of efficiency analysis in the economics industrial organization and regulation is the determination of efficiency or production frontiers (Timmer 1971). Given cost of production and possibly other factors, X, we are interested in the highest production levels that only a small fraction of firms, the most efficient firms, can attain. These (nearly) efficient production levels can be formally described by extremal quantile regression function, $F_Y^{-1}(\tau|X)$ for $\tau \in [1-\varepsilon, 1]$ and $\varepsilon > 0$; so only an ε-fraction of firms produce $F_Y^{-1}(\tau|X)$ or more. The models and methods discussed in this article are highly pertinent for inference on the probabilistic frontiers.

Example 4 (S,s)-Rules and Other Approximate Reservation Rules in Economic Decisions A related example is that of (S,s)-adjustment models, which arise as optimal policies in many economic models (Arrow et al. 1951). For example, the capital stock Z is adjusted up to the level S once it has depreciated to some low level s. In terms of an econometric specification, we may think that the observed capital stock satisfies the equation $Z_i = s(X_i) + v_i$, where X_i are covariates, and v_i is a disturbance that is positive most of the time, that is, Prob $(v_i \geq 0)$ is close to 1. Once the capital stock Z_i reaches the critical level, that is $Z_i \leq s(X_i)$, it is adjusted in the next period. We assume, as in Caballero and Engel (1999), that when the disturbance $v_i = Z_i - s(X_i)$ is negative, it captures unobserved heterogeneity and small decision mistakes that are independent of observed covariates X_i. (In this example, Z_i

could be any monotone transformation of the stock variable. For instance, the log transformation gives an accelerated failure time model for the capital stock. Caballero and Engel 1999, explore such specifications for empirical (S,s) models in detail). In a given cross-section or time series, adjustment will occur infrequently, so in fact data at or below the lower adjustment band $s(X_i)$ will be observed with a small probability Prob $(v_i \leq 0)$; hence $F_Z^{-1}(\tau|X) = s(X) + F_v^{-1}(\tau)$ for $\tau \in (0, \text{Prob}\ (v \leq 0))$. The lower-band function $s(X)$ therefore coincides with the lower conditional quantile function up to an additive constant. A similar argument works for the upper-band function $S(X)$.

Example 5 Structural Auction Models In the standard specification of the first-price procurement auction where bidders hold independent valuations, the winning bid, B_i, satisfies the equation $B_i = c(X_i)\beta(n_i))\ \varepsilon_i$, $\varepsilon_i \geq 0$, where $c(X_i)$ is the efficient cost function and $\beta(n_i) \geq 1$ is a mark-up that approaches 1 as the number of bidders, n_i, approaches infinity (Donald and Paarsch 2002). By construction, $c(X_i)\beta(n_i)$ is the extreme conditional quantile function. In empirical analysis, it is realistic to let the disturbance ε_i take some small negative value so that, when negative, these disturbances capture small decision mistakes that are independent of included explanatory variables. In this case the quantile function satisfies $F_B^{-1}(\tau|X, n) = C(X)\beta(n) + F_\varepsilon^{-1}(\tau)$, for $\tau \in (0, P[\varepsilon \leq 0])$. Quantile regression methods can be employed to make inference on $c(X)$ and $\beta(n)$.

Organization of the Article

The rest of the article is organized as follows. Section "Basic Models of Extremal Quantiles" describes the basic model of extremal quantiles and extremal conditional quantiles. Section "Basic Estimation Methods" describes basic estimation theory and inference theory. Section "Empirical Applications: An Overview and an Illustration" reviews the key empirical applications and provides an illustrative example. Section "Conclusion" concludes.

Basic Models of Extremal Quantiles

A Basic Model of Extremal Quantiles

Towards discussing inference methods, assume that the distribution function of the response variable Y has Pareto-type tails, which means that tails behave approximately like power functions. Such tails are prevalent in economic data, as discovered by the prominent Italian econonometrician Vilfredo Pareto in 1895 (see Pareto 1964). Pareto-type tails encompass or approximate a rich variety of tail behaviour, including that of thick-tailed and thin-tailed distributions, having either bounded or unbounded support, and their mathematical theory in connection to extreme value theory has been developed by Gnedenko (1943) and de Haan (1970).

Consider a random variable Y and define a random variable U as $U \equiv Y$, if the lower end-point of the support of Y is $-\infty$, and $U \equiv Y - F_Y^{-1}(0)$, if the lower end-point of the support of Y is $F_Y^{-1}(0) > -\infty$. The distribution function of U, denoted by F_U then has the lower end-point $F_U^{-1}(0) > -\infty$ or $F_U^{-1}(0) = 0$. The assumption that the distribution function F_U and its quantile function F_U^{-1} exhibit Pareto-type behaviour in the tails can be formally stated as the following two equivalent conditions (the notation $a \sim b$ means that $a/b \to 1$ as appropriate limits are taken):

$$F_U(u) \sim \overline{L}(u) \cdot u^{-1/\xi} \text{ as } u \searrow F_U^{-1}(0), \qquad (1)$$

$$F_U^{-1}(\tau) \sim L(\tau) \cdot \tau^{-\xi} \text{ as } \tau \searrow 0, \qquad (2)$$

for some real number $\xi \neq 0$, where $\overline{L}(u)$ is a nonparametric, slowly varying function at $F^{-1}(0)$, and $L(\tau)$ is a nonparametric slowly varying function at 0 (A function $u \mapsto L(u)$ is said to be slowly varying at 0 if $\lim_{l \searrow s}[L(l)/L(ml)] = 1$ for any $m > 0$). The prime examples of slowly varying functions are the constant function $L(y) = L$ and the logarithmic function. The number ξ defined in (1) and (2) is called the extreme value (EV) index.

The absolute value $|\xi|$ of the EV index ξ measures heavy tailedness of distributions. A distribution F_Y with Pareto-type tails

necessarily has a finite lower support point if $\xi < 0$ and an infinite lower support point if $\xi > 0$. Distributions with $\xi > 0$ include stable distributions, Pareto distributions, t distributions, and many others. For example, the t distribution with v degrees of freedom has the EV index $\xi = 1/v$ and exhibits a wide range of tail behaviour. In particular, setting $v = 1$ yields the Cauchy distribution which has heavy tails (with $\xi = 1$), while setting $v = 30$ yields approximately the normal distribution which has light tails (with $\xi = 1/30$). On the other hand, distributions with $\xi < 0$ include the uniform, exponential, Weibull distributions, and others.

The assumption of Pareto-type tails can be equivalently cast in terms of the regular variation assumption, as is commonly done in EV theory. Distribution function F_U is said to be *regularly varying* at $F_U^{-1}(0)$ with index of regular variation $-1/\xi$ if $\lim_{y \searrow F_U^{-1}(0)} (F_U(ym)/F_U(y) = m^{-1/\xi}$, for any $m > 0$. This condition is equivalent to the regular variation of quantile function F_U^{-1} at 0 with index $-\xi$:

$$\lim_{\tau \searrow 0} (F_U^{-1}(\tau m)/F_U^{-1}(\tau) = m^{-\xi}, \text{for any } m > 0.$$

It should be mentioned that the case of $\xi = 0$ corresponds to the class of rapidly varying distribution functions. These distribution functions have exponentially light tails, with the normal and exponential distributions being the chief examples. To simplify exposition, we do not discuss this case explicitly. However, since the limit distribution of main statistics are continuous in ξ, including at $\xi = 0$, inference theory for the case of $\xi = 0$ can be adequately approximated by the case of $\xi \approx 0$.

A Basic Model of Extremal Conditional Quantiles

Consider the classical linear functional form for the conditional quantile function of Y given $X = x$:

$$\begin{aligned} F_U^{-1}(\tau | x) &= x' \beta(\tau), \text{for all } \tau \in \mathcal{I} \\ &= (0, \eta], \text{some } \eta \in (0, 1], \end{aligned} \tag{3}$$

and for every x in the support of X. This linear functional form is flexible in the sense that it has good approximation properties. Given the original regressor X^*, the final set of regressors X to be used in estimation can be formed as a vector of approximating functions. For example, X may include power functions, splines, and other transformations of X^*. The linear functional form also provides computational convenience.

The following model for the tails and its generalizations were developed in Chernozhukov (2005). The main assumption is that the response variable Y, transformed by some auxiliary regression line, has regularly varying tails with EV index ξ. Indeed, in addition to (3), suppose there exists an auxiliary parameter β_e such that the disturbance $U \equiv Y - X'\beta_e$ has conditional end-point 0 or $-\infty$ a.s. and its conditional quantile function $F_U^{-1}(\tau | x)$ satisfies the following tail-equivalence relationship as $\tau \searrow 0$, uniformly for x in the support of X:

$$F_U^{-1}(\tau | x) = F_Y^{-1}(\tau | x) - x' \beta_e \sim F_u^{-1}(\tau), \tag{4}$$

where F_u^{-1} is a quantile function such that

$$F_u^{-1}(\tau) \sim L(\tau) \tau^{-\xi}, \tag{5}$$

where $L(\tau)$ is a non-parametric slowly varying function at 0. Equation (5) imposes Pareto-type behaviour on the conditional law, while Eq. (4) requires this behaviour to hold uniformly across conditioning values. Since this assumption only affects the tails, it allows covariates to impact the extremal quantiles and the central quantiles very differently; the impact of covariates on extremal quantiles is approximated by β_e, which could differ sharply from, for example, the impact on the median given by $\beta(1/2)$. Chernozhukov (2005) provides further generalizations of this model.

Basic Estimation Methods

Estimates Based on Sample Quantiles

Given T observations $\{Y_t, t = 1, \ldots, T\}$, the τ-sample quantile can be obtained by solving the following optimization problem:

$$\hat{F}_Y^{-1}(\tau) \in \arg\min_{\beta \in R^d} \sum_{t=1}^{T} \rho_\tau(Y_t - \beta), \qquad (6)$$

where $\rho_\tau(u) = (\tau - 1(u < 0))u$ is the asymmetric absolute deviation function of Fox and Rubin (1964).

Sample quantiles are also order statistics, and we will refer to τT as the *order* of τ-quantile. The sequence of quantile index-sample size pairs (τ, T) will be said to be an *extreme order* sequence if $\tau \searrow 0$ and $\tau T \to k > 0$, an *intermediate order* sequence if $\tau \searrow 0$ and $\tau T \to \infty$, and a *central order* sequence if τ is fixed and $T \to \infty$. Each different type of the sequence leads to different asymptotic approximations to the finite-sample distributions of sample quantiles. Extreme order sequences lead to non-normal (extreme-value) distributions (EV) that approximate the finite sample distributions of extremal (high and low) quantiles much better than the normal distributions do. In particular, EV distributions work much better than the normal distribution if $\tau T \leq 30$.

Extreme Order Quantiles

Consider an extreme-order sequence. The following is the classical result on the limit distribution of order statistics: for any integer $k \geq 1$ and $\tau = k/T$, as $T \to \infty$,

$$A_T\left(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau)\right) \to {}_d\Gamma_k^{-\xi} - k^{-\xi} \qquad (7)$$

where

$$A_T = 1/F_U^{-1}(1/T), \quad \Gamma_k = \mathscr{E}_1 + \cdots + \mathscr{E}_k \qquad (8)$$

and $(\mathscr{E}_1, \mathscr{E}_2, \ldots)$ is an independent and identically distributed sequence of standard exponential variables.

Result (7) was obtained by Gnedenko (1943) under the assumption that $Y_1, Y_2, \ldots$ is a sequence of independently and identically distributed (i.i.d.) random variables. Result (7) continues to hold for stationary weakly dependent series, provided the probability of extreme events occurring in clusters is negligible relative to the probability of a single extreme event (Meyer 1973). The results have been generalized to more general time series processes (Leadbetter et al. 1983).

Result (7) gives an EV distribution as an approximation to the finite-sample distribution of $\hat{F}_Y^{-1}(\tau)$ The EV distribution is characterized by the EV index ξ, which can be estimated by one of the methods described below. Variables Γ_k, entering the definition of the EV distribution, are known as *gamma* random variables. The limit distribution of the kth-order statistic is therefore a transformation of a gamma variable. The EV distribution is not symmetric and may have significant (median) bias. The EV distribution has finite moments if $\xi < 0$ and has finite moments of up to order $1/\xi$ if $\xi > 0$.

The classical result is not feasible for purposes of inference on $F_Y^{-1}(\tau)$, since the scaling constant A_T is not easily estimable consistently. One way to overcome this problem is to make additional strong assumptions in order to estimate A_T consistently. For instance, suppose that $F_U^{-1}(\tau) \sim L\tau^{-\xi}$, then one can estimate ξ using methods described below and L by $\hat{L} = \left(\hat{F}_Y^{-1}(2\tau) - \hat{F}_Y^{-1}(\tau)\right) / \left(2^{-\hat{\xi}} - 1\right)\tau^{-\hat{\xi}}$.

Another way to overcome the aforementioned infeasibility is to consider the asymptotics of self-normalized extreme order quantiles, as in Chernozhukov (2006):

$$Z_T(\tau) = \mathcal{A}_T\left(\hat{F}_Y^{-1}(\tau) - F_Y^{-1}(\tau)\right) \to d\frac{\sqrt{k}\left(\Gamma_k^{-\xi} - k^{-\xi}\right)}{\Gamma_{mk}^{-\xi} - \Gamma_k^{-\xi}} \qquad (9)$$

where for $m > 1$ such that mk is an integer,

$$\mathcal{A}_T = \frac{\sqrt{\tau T}}{\hat{F}_Y^{-1}(m\tau) - \hat{F}_Y^{-1}(\tau)}. \qquad (10)$$

Here, the scaling factor A_T is feasible in that it is completely a function of data. The limit distribution only depends on the EV index ξ, and its quantiles can be easily calculated analytically or by simulation.

E

Intermediate Order Quantiles

Consider next an intermediate order sequence. As $\tau \searrow 0$ and $\tau T \to \infty$, under further regularity conditions,

$$Z_T(\tau) = \mathcal{A}_T\left(F_Y^{-1}(\tau) - F_Y^{-1}(\tau)\right) \to_d \mathcal{N}\left(0, \frac{\xi^2}{(m^\xi - 1)^2}\right), \qquad (11)$$

where A_T is defined as in (10). This result, obtained by Dekkers and de Haan (1989), gives a normal asymptotic approximation to the finite-sample distribution of sample quantile $F_Y^{-1}(\tau)$ The main condition for application of this distribution is that $\tau T \to \infty$. In finite samples, we may interpret this as requiring that $\tau T \geq 30$ at the minimum.

The normal approximation (11) is convenient, but extreme approximation (9) is always better, because it does not fail when $\tau T \to k < \infty$ and it coincides with the normal approximation (11) once k is large.

Extremal Bootstrap

In many cases it is convenient to implement inference using the following approach. Consider the sample of i.i.d. variables:

$$(Y_1, \ldots, Y_T) = \left(\frac{\mathscr{E}_1^{-\xi} - 1}{-\xi}, \ldots, \frac{\mathscr{E}_T^{-\xi} - 1}{-\xi}\right), \qquad (12)$$

where $(\mathscr{E}_1, \ldots, \mathscr{E}_T)$ is an i.i.d. sequence of standard exponential variables. (Y_t, defined in this way, follows generalized extreme value distribution, which nests the Frechet, Weibull, and Gumbell distributions. There are other possibilities, for example, $(\mathscr{E}_1, \ldots, \mathscr{E}_T)$ in (13) can be replaced by uniform variables $(U_1, \ldots, U_T)$, in which case Y_t follows the generalized Pareto distribution.) Variables generated in this way have the quantile function:

$$F_Y^{-1}(\tau) = \frac{[-\ln(1-\tau)]^{-\tau} - 1}{-\xi} \qquad (13)$$

Observe that $F_Y^{-1}(\tau) - 1/\xi \sim \tau^{-\xi}/\xi$, so condition (1) is satisfied. We propose to estimate the finite-sample distributions of $Z_T(\tau) = \mathscr{A}_T\left(F_Y^{-1}(\tau) - F_Y^{-1}(\tau)\right)$ by the finite-sample distribution of $Z_T(\tau)$ for the case when the data follow (13). In this way, we reproduce both the EV limit (9) and the normal limit (11) under extreme and intermediate sequences, and also guarantee good finite-sample performance for the case when (13) holds exactly. The simulation can be done using the following algorithm:

1. For each $i \leq B$, draw $(Y_1, \ldots, Y_T)$ as i.i.d. according to (13), replacing ξ with a suitable estimate $\hat{\xi}$. Compute the statistic $Z_{T,i}(\tau) = \mathcal{A}_T\left(F_Y^{-1}(\tau) - F_Y^{-1}(\tau)\right)$.
2. Use quantiles of the simulated sample $(Z_{T,i}(\tau), i \leq B)$ for inference purposes.

This scheme could be used to estimate distributions of other statistics, including estimators of the EV index and extrapolation estimators.

Another method, developed in Chernozhukov (2006), is based on subsampling the self-normalized quantile statistic. This method is less accurate than the extremal bootstrap. However, it applies under more general conditions. It should be noted that the canonical (nonparametric) bootstrap does not work in these settings (Bickel and Freedman 1981).

Confidence Intervals for $F_Y^{-1}(\tau)$ and Bias Correction for $F_Y^{-1}(\tau)$

Let the α-quantile of $Z_T(\tau)$ be denoted by $c(\alpha)$. The estimates of $c(\alpha)$ can be obtained using either EV approximation, normal approximation, or the extremal bootstrap, also having replaced ξ with a suitable estimate. Denote the resulting estimates by $\hat{c}(\alpha)$. Then, the median bias-corrected estimate and α%-confidence region for $F_Y^{-1}(\tau)$ can be constructed as

$$F_Y^{-1}(\tau) - \frac{\hat{c}(1/2)}{\mathcal{A}_T} \quad \text{and} \quad \left[\hat{F}_Y^{-1}(\tau) - \frac{\hat{c}(1-\alpha/2)}{\mathcal{A}_T}, \hat{F}_Y^{-1}(\tau) - \frac{\hat{c}(\alpha/2)}{\mathcal{A}_T}\right]. \qquad (14)$$

Estimators of the EV Index ξ

There are two principal estimators. The first estimator, due to Pickands (1975), relies on the ratio of sample quantile spacings:

$$\hat{\xi} = -\ln\left[\frac{\hat{F}_Y^{-1}(4\tau) - \hat{F}_Y^{-1}(2\tau)}{\hat{F}_Y^{-1}(2\tau) - \hat{F}_Y^{-1}(\tau)}\right]/\ln 2, \qquad (15)$$

such that $\tau \to 0$ and $\tau T \to \infty$ as $T \to \infty$. Under further regularity conditions,

$$\sqrt{\tau T}\left(\hat{\xi} - \xi\right) \to_d \mathcal{N}^d\left(0, \frac{\xi^2\left(2^{2\xi+1}+1\right)}{\left(2\left(2^{\xi}-1\right)\ln 2\right)^2}\right) \qquad (16)$$

Another estimator, developed by Hill (1975), is a moments estimator (notation $(x)_-$ means $(x)_- = -x$ if $x < 0$ and $(x)_- = 0$ if $x \geq 0$):

$$\hat{\xi} = \frac{\sum_{t=1}^{T} \ln\left(Y_t/\hat{F}_Y^{-1}(\tau)\right)_-}{T\tau} \qquad (17)$$

such that $\tau \to 0$ and $\tau T \to \infty$ as $T \to \infty$. This estimator is applicable only for the case of $\xi > 0$. The estimator can be motivated by a maximum likelihood method that fits an exact power law to the tail data. Under further regularity conditions,

$$\sqrt{\tau T}\left(\hat{\xi} - \xi\right) \to_d \mathcal{N}\left(0, \xi^2\right). \qquad (18)$$

The methods for choosing τ are described in Embrechts et al. (1997). The variance of estimators decreases as τ increases, but the bias (relative to the true ξ) goes up. Another view on the choice of τ is the following: statistical models are approximations, not literal descriptions of the data. In practice, dependence of $\hat{\xi}$ on the threshold τ reflects that power laws with different values of ξ fit better different tail regions. Therefore, if the interest lies in making inference on $\hat{F}_Y^{-1}(\tau)$ for a particular τ, it seems reasonable to use ξ constructed using the same τ or most similar τ' subject to the condition that $\tau' T \geq 30$ (The latter condition requires that a sufficient sample be available to estimate ξ).

The limit results above can be used for the construction of confidence regions. As an alternative, we can apply extremal bootstrap to statistic $Z_T = \sqrt{\tau T}\left(\hat{\xi} - \xi\right)$. To estimate the quantiles of Z_T. Given the estimated α-quantiles $\hat{c}(\alpha)$, we can construct the median bias-corrected estimate and α%-confidence regions for ξ.

Extrapolation Estimators

When very extreme quantiles cannot be estimated precisely, the following strategy is sensible: estimate less extreme quantiles reliably, and then extrapolate these estimates using the assumptions on tail behaviour stated earlier. Dekkers and de Haan (1989) developed the following extrapolation estimator:

$$\hat{F}_Y^{-1}(\tau_e) = \frac{(\tau_e/\tau)^{-\hat{\xi}} - 1}{2^{-\hat{\xi}} - 1}\left[\hat{F}_Y^{-1}(2\tau) - \hat{F}_Y^{-1}(\tau)\right] + \hat{F}_Y^{-1}(\tau), \qquad (19)$$

where $\tau_e \ll \tau$. Another useful estimator, which is valid only for the case of $\xi > 0$, is the following:

$$\hat{F}_Y^{-1}(\tau_e) = (\tau_e/\tau)^{-\hat{\xi}} \cdot \hat{F}_Y^{-1}(\tau), \qquad (20)$$

where $\tau_e \ll \tau$. The above estimators have good properties provided the quantities on the right-hand side of (19) and (20) are well estimated, which requires that τT be large, and that the tail model be a good approximation of the underlying true tail.

Estimates Based on Sample Regression Quantiles

Given T observations $\{Y_t, X_t, t = 1, \ldots, T\}$, the quantile regression estimate of $F_Y^{-1}(\tau|x)$ is given by:

$$\hat{F}_Y^{-1}(\tau|x) = x'\hat{\beta}(\tau), \hat{\beta}(\tau) = \arg\min_{\beta \in \mathbb{R}^d} \sum_{t=1}^{T} \rho_\tau\left(Y_t - X_t'\beta\right), \qquad (21)$$

where $\rho_\tau(u) = (\tau - 1(u < 0))u$. Quantile regression was introduced by Laplace (1818) for the

median case. Koenker and Bassett (1978) extended this formulation to other quantiles.

Extreme Order Asymptotics

Chernozhukov (2005) derives asymptotic distributions of regression quantiles under extreme-order sequences. Consider the canonically normalized QR statistic

$$\hat{Z}_T(k) = A_T\left(\hat{\beta}(\tau) - \beta(\tau)\right), \text{where } A_T = 1/F_u^{-1}(1/T), \tag{22}$$

and the self-normalized QR statistic

$$Z_T(k) = \mathcal{A}_T\left(\hat{\beta}(\tau) - \beta(\tau)\right), \text{where } \mathcal{A}_T = \frac{\sqrt{\tau T}}{\overline{X}'\left(\hat{\beta}(m\tau) - \hat{\beta}(\tau)\right)}, \tag{23}$$

where $\tau T(m-1) > d$. The first statistic uses an infeasible canonical normalization, while the second statistic uses a feasible normalization. Then as $\tau T \to k > 0$ and $T \to \infty$

$$\begin{aligned}\hat{Z}_T(\tau) &\to_d Z^d\infty(k) - k^{-\xi}, Z_\infty(k) \\ &= \text{argmin}_{z\in\mathbb{R}^d}\left[-kE[X]'z + \sum_{i=1}^{\infty}\left[X_i'z - \Gamma_i^{-\xi}\right]_+\right] \\ &\quad \times (\xi < 0)\, Z_\infty(k) = \text{argmin}_{z\in\mathbb{R}^d} \\ &\left[-kE[X]'z + \sum_{i=1}^{\infty}\left[X_i'z - \Gamma_i^{-\xi}\right]_+\right] \\ &\quad \times (\xi > 0)\end{aligned} \tag{24}$$

where $\{\Gamma_1, \Gamma_2, \ldots\} := \{\mathscr{E}_1, \mathscr{E}_1 + \mathscr{E}_2, \ldots\}$ and $\{\mathscr{E}_1, \mathscr{E}_1, \ldots\}$ is an i.i.d. sequence of exponential variables that is independent of $\{X_1, X_2, \ldots\}$. Further, for any m such that $k(m-1) > d$,

$$Z_T(\tau) \to_d \frac{\sqrt{k}Z_\infty(k)}{E[X]'(Z_\infty(mk) - Z_\infty(k))}. \tag{25}$$

The results hold under the assumption that the data come from either an i.i.d. sequence or a stationary weakly dependent sequence with extreme events satisfying a non-clustering condition.

Related results for canonically normalized statistics for the case where $\tau T \to 0$ as $T \to \infty$ have been obtained by Knight (2001) and Portnoy and Jurečková (1999).

Intermediate Order Asymptotics

Chernozhukov (2005) shows that under intermediate order sequences, as $\tau \searrow 0$ and $\tau T \to \infty$,

$$\begin{aligned}Z_T(\tau) &= \mathcal{A}_T\left(\hat{\beta}(\tau) - \beta(\tau)\right) \\ &\to_d \mathcal{N}^d\left(0, [E(XX')]^{-1}\frac{\xi^2}{(m^{-\xi}-1)^2}\right),\end{aligned} \tag{26}$$

where $\mathscr{A}_T$ is defined as in (28). Like the result under extreme order sequences, this result holds under the assumption that the data come either from an i.i.d. sequence or from a stationary weakly dependent sequence with extreme events satisfying a non-clustering condition.

Extremal Bootstrap

In practice, it is convenient to implement inference by constructing a bootstrap model that approximates the tail features of the true conditional quantile model under the assumptions of Sect. "A Basic Model of Extremal Conditional Quantiles"; then using this model to simulate the distributions of estimators of extreme quantiles and tail parameters.

Consider the sample

$$\begin{aligned}&((Y_1, X_1), \ldots, (Y_T, X_T)) \\ &= \left(\left(\frac{\mathscr{E}_1^{-\xi} - 1}{-\xi}, X_1\right), \ldots, \left(\frac{\mathscr{E}_T^{-\xi} - 1}{-\xi}, X_T\right)\right),\end{aligned} \tag{27}$$

where $(\mathscr{E}_1, \ldots, \mathscr{E}_T)$ is an i.i.d. sequence of standard exponential variables and $(X_1, \ldots, X_T)$ is a fixed set of observations on regressors that we have. Variable Y_t generated in this way has the conditional quantile function

$$\begin{aligned}&F_{Y_t}^{-1}(\tau|X_t) = X_t'\beta(\tau) = \frac{[-\ln(1-\tau)]^{-\xi} - 1}{-\xi}, \\ &\text{where } \beta(\tau) = \left(\frac{[-\ln(1-\tau)]^{-\xi} - 1}{-\xi}, 0, \ldots, 0\right)'.\end{aligned} \tag{28}$$

Observe that $F_{Y_t}^{-1}(\tau|X_t) - 1/\xi \sim \tau^{-\xi}/\xi$, so the model satisfies conditions (4) and (5), as does the true conditional quantile model under our assumptions. Hence we can estimate the finite-sample distributions of $Z_T(\tau) = \mathcal{A}_T\left(\hat{\beta}(\tau) - \beta(\tau)\right)$ by the finite-sample distribution of $Z_T(\tau)$ in the case when data follows (32). In this simple way, we can replicate both the EV approximation (31) and the normal approximation (24) and also guarantee good finite-sample performance for the case when the model (32) holds. The simulation can be done using the following algorithm:

1. For each $i \leq B$, draw data according to (32), replacing ξ with a suitable estimate $\hat{\xi}$. Compute the statistic $Z_{T,i}(\tau) = \mathcal{A}_T\left(\hat{\beta}(\tau) - \beta(\tau)\right)$.
2. Use the empirical distribution of the simulated sample $(Z_{T,i}(\tau), i \leq B)$ for inference.

This method can also be used to estimate distributions of estimators of the EV index and extrapolation estimators described below.

As mentioned before, there is another inference method proposed by Chernozhukov (2006) which uses subsampling to estimate the distribution of the self-normalized statistic $Z_T(\tau)$. This method is less accurate than the extremal bootstrap, but it applies under more general conditions.

Confidence Intervals and Bias Corrected Estimates

Suppose we are interested in the parameter $\psi'\beta(\tau)$ for some non-zero vector ψ. Let the α-quantile of $\psi' Z_T(\tau)$ be denoted by $c(\alpha)$. Having replaced ξ with a suitable estimate, the estimates of $c(\alpha)$ can be obtained using either the EV approximation, normal approximation or the extremal bootstrap. Denote the resulting estimates by $\hat{c}(\alpha)$. The median-bias corrected estimator and the α%-confidence interval for $\psi'\beta(\tau)$ can be constructed as

$$\psi'\hat{\beta}(\tau) - \frac{\hat{c}(1/2)}{\mathcal{A}_T} \text{ and } \left[\psi'\hat{\beta}(\tau) - \frac{\hat{c}(\alpha/2)}{\mathcal{A}_T}, \psi'\hat{\beta}(\tau) - \frac{\hat{c}(1-\alpha/2)}{\mathcal{A}_T}\right]. \tag{29}$$

Estimators of the EV Index ξ

The following estimators are regression analogs of the Pickands and Hill estimators. The first estimator takes the form

$$\hat{\xi} = -\ln\left[\frac{\hat{F}_Y^{-1}(4\tau\overline{X}) - \hat{F}_Y^{-1}(2\tau\overline{X})}{\hat{F}_Y^{-1}(2\tau\overline{X}) - \hat{F}_Y^{-1}(\tau\overline{X})}\right]/\ln 2, \tag{30}$$

where $\overline{X}$ is the average value of X_t. Under additional regularity conditions, as $\tau \searrow 0$ and $\tau T \to \infty$

$$\sqrt{\tau T}\left(\hat{\xi} - \xi\right) \to_d \mathcal{N}\left(0, \frac{\xi^2\left(2^{2\xi+1}+1\right)}{\left(2\left(2^{\xi}-1\right)\ln 2\right)^2}\right). \tag{31}$$

The second estimator, which is applicable when $\xi > 0$, takes the form:

$$\hat{\xi} = -\frac{\sum_{t=1}^{T} \ln\left(Y_t/\hat{F}_Y^{-1}(\tau|X_t)\right)}{T\tau}. \tag{32}$$

Under additional regularity conditions, as $\tau \searrow 0$ and $\tau T \to \infty$

$$\sqrt{\tau T}\left(\hat{\xi} - \xi\right) \to_d \mathcal{N}\left(0, \xi^2\right). \tag{33}$$

The limit results above can be used for the construction of confidence regions. An alternative approach is to apply the extremal bootstrap to statistic $Z_T = \sqrt{\tau T}\left(\hat{\xi} - \xi\right)$ to estimate the quantiles of this statistic. Then we can use estimated α-quantiles $\hat{c}(\alpha)$ for constructing the median bias-corrected estimate and α%-confidence regions for ξ:

$$\hat{\xi} - \frac{\hat{c}(1/2)}{\sqrt{\tau T}} \text{ and } \left[\hat{\xi} - \frac{\hat{c}(1-\alpha/2)}{\sqrt{\tau T}}, \hat{\xi} - \frac{\hat{c}(\alpha/2)}{\sqrt{\tau T}}\right]. \tag{34}$$

Extrapolation Estimators

By analogy with the unconditional case, the extrapolation estimators for $\hat{F}_Y^{-1}(\tau_e\overline{X})$, where τ_e is a very low value, can be constructed as

$$\hat{F}_Y^{-1}(\tau_e|x) = \frac{(\tau_e/\tau)^{-\hat{\xi}} - 1}{m^{-\hat{\xi}} - 1} \left[\hat{F}_Y^{-1}(m\tau|x) - \hat{F}_Y^{-1}(\tau|x)\right] + \hat{F}_Y^{-1}(\tau|x), \quad (35)$$

$$\hat{F}_Y^{-1}(\tau_e|x) = (\tau_e/\tau)^{-\hat{\xi}} \hat{F}_Y^{-1}(\tau|x)(\xi > 0), \quad (36)$$

where $\tau_e \ll \tau$. Note that the estimator (41) is valid only in the case $\xi > 0$. The comments given for the unconditional case apply here as well. Also, we can construct confidence regions for $F_Y^{-1}(\tau_e|x)$ based on extrapolation estimators. This can be done by applying the extremal bootstrap to statistic $Z_T = \mathcal{A}_T\left(\hat{F}_Y^{-1}(\tau_e|x) - \hat{F}_Y^{-1}(\tau_e|x)\right)$ where $\mathcal{A}_T = \sqrt{\tau T}/\overline{X}'\left(\hat{\beta}(2\tau) - \hat{\beta}(\tau)\right)$.

Empirical Applications: An Overview and an Illustration

A Simple Overview

The following review is not exhaustive by any means; it aims to provide only a few quintessential references.

Extremal Unconditional Quantiles

As mentioned in Sect. "Basic Models of Extremal Quantiles", Pareto analysed income and wealth data in 1895 and suggested that power laws accurately describe the tail data. Pareto's discovery, although rmkably simple, had a profound effect on both empirics and the theory of extremes. Zipf (1949), Mandelbrot (1963), Fama (1965), Praetz (1972), Sen (1973), Jansen and de Vries (1991), and Longin (1996), among others, gave further empirical evidence on the nature and prevalence of Pareto-type laws in economic data, including city sizes, incomes, and financial returns.

It should be mentioned that many of the early studies were highly informal in nature. The theoretical work in extreme value theory has opened paths for better analysis. From this aspect, the study of Jansen and de Vries (1991) can be singled out as it gave, to our knowledge, the first highly rigorous analysis of the tail properties of financial returns. Jansen and de Vries (1991) estimate the EV indices for various primary US stocks to be between $\xi = 1/5$ and $\xi = 1/3$. Using quantile extrapolation estimators to estimate value-at-risk, they also conclude that the 1987 market crash was not an outlier. Rather it was a rare event, the magnitude of which could have been predicted using prior data. This study stimulated numerous other studies that rigorously document the tail properties of economic data (Embrechts et al. 1997).

Extremal Conditional Quantiles

There has been considerably less work on conditional methods. However, following recent theoretical advances we expect that this area will see active development in the near future. In what follows, we merely highlight some of the topics and directions.

In what might be the earliest example of conditional quantile analysis, Quetelet (1871) fitted various conditional quantile curves to age–height data. Remarkably, Quetelet's work included tabulations of very high and very low quantiles of heights as a function of age. There is a great potential for the applications of extremal quantile regression methods in similar problems. In a recent study, Chernozhukov (2006) estimates the impact of smoking and maternal behaviour on extremely low birthweights in the United States, focusing on black mothers. He finds that the impact of these variables on birthweights in the ranges between 250 and 1500 g sharply differs from their impact on the central birthweights. For instance, smoking is not correlated with extremal birthweights, while quality of prenatal medical care is strongly linked to extremal birthweights.

Aigner and Chu (1968), Timmer (1971), and Aigner et al. (1976) pioneered a large empirical literature on production frontiers. A major problem of the subsequent empirical literature has been the lack of statistical methods for construction of reliable estimates and confidence regions. The new methods discussed in Sect. "Basic Estimation Methods" solve the problem, and should improve the rigour of the empirical work in this area.

There is a considerable appeal for the use of extremal conditional quantile methods in auction models. An important study that illustrates the potential is by Donald and Paarsch (2002), who analyse an empirical structural auction model.

They estimate the conditional support function of bids using extreme order statistics for each covariate cell, then project the estimated function onto a lower dimensional structural function implied by the model via a minimum-distance method. A generalization of this approach is to employ extremal quantile regression for estimation of the (approximate) support function in the first stage (The use of near-extreme quantile regression for estimation of approximate support functions allows the researcher to discard some outliers that do not conform the model, in the spirit of the discussion given in Sect. "Introduction").

Value-at-risk is another potentially important area of applications of the extremal quantile regression. Chernozhukov and Umantsev (2001) apply these methods to the problem of forecasting value-at-risk of a major US oil company. They estimate extremal conditional quantiles, using both ordinary and extrapolation methods, and implement confidence regions, using subsampling methods described in Chernozhukov (2006). The section below briefly revisits some of the main qsts of this study.

An Illustrative Example

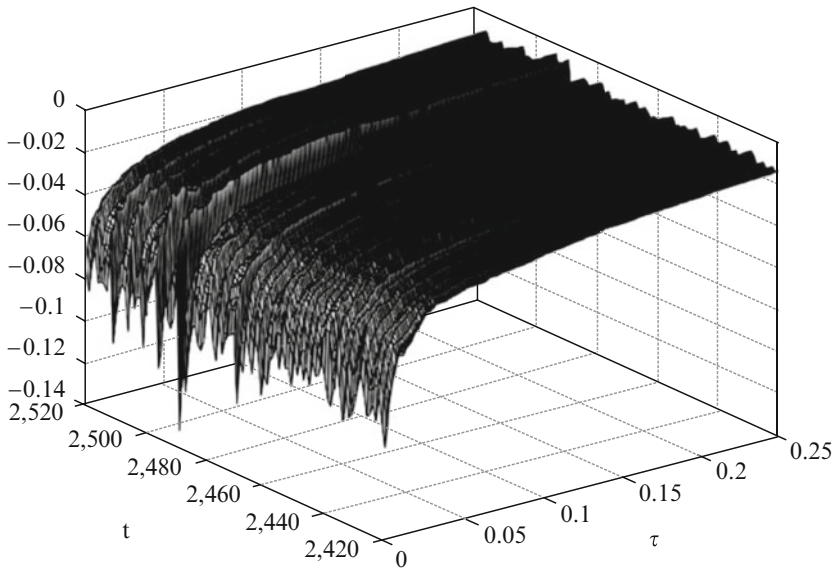
Here we consider a problem of forecasting conditional value-at-risk. We revisit some of the qsts asked in Chernozhukov and Umantsev (2001) with an improved methodology. To implement the analysis, we use algorithms written in R language that rely on Koenker's (2006) quantreg package as the basic platform. The algorithms as well as the data-set can be downloaded from http://www.mit.edu/~vchern/EQR (accessed 23 April 2007). A detailed description of the data-set is given in Chernozhukov and Umantsev (2001).

We estimate the following conditional quantile function for various low values of τ:

$$\begin{aligned}\hat{F}_Y^{-1}(\tau|X_t) &= X_t'\beta(\tau)\\ &= \beta_0(\tau)+\beta_1(\tau)X_{t,1}\\ &\quad +\beta_2(\tau)X_{t,2}+\beta_3(\tau)X_{t,3},\end{aligned} \quad (37)$$

where Y_t is the daily (log) return on the stock of Occidental Petroleum, $X_{t,1}$ is the lagged return on spot oil price, $X_{t,2}=Y_{t-1}$ is the lagged own return, and $X_{t,3}$ is the lagged return on the Dow Jones Industrial Index. There are 2527 observations in the sample.

We first estimate and plot the function $(\tau,t)\mapsto X_t'\beta(\tau)$ in (τ,t) space, with $\tau\le 0.25$. The graph of this function, shown in Fig. 1, gives a good picture of the evolution of risk over time, indicating dates



Extremal Quantiles and Value-at-Risk, Fig. 1 The fit $X_t'\beta(\tau)$ as a function of time t and τ (*Source*: Chernozhukov and Umantsev (2001))

where the predicted risk is especially high. Let us next determine what causes these risk fluctuations.

Table 1 reports the estimates of the coefficients of the model (37). The table also reports median bias-corrected estimates and 90% confidence regions, which were obtained using the extremal bootstrap approach described in Sect. "Basic Estimation Methods". The results show that the primary determinant of the high-risk levels is the market. The further in the tail we go, the larger is the magnitude of the point estimate of the coefficient on the DJI return. Moreover, the confidence region for this coefficient excludes zero even at $\tau = 0.001$.

We next characterize the tail properties of the conditional quantile model. Table 2 reports the estimates of the EV index ξ obtained using estimator (37). The table also reports median bias-corrected estimates, and 90% confidence regions, which were obtained using the nested approach described in Sects. "Extremal Bootstrap" and "Estimators of the EV Index ξ". The bias-corrected estimates tend to be stable with respect to the start of the tail determined by probability index τ. On the basis of Table 2, we take $\hat{\xi} \approx 1/4$ to be the estimate of the EV index.

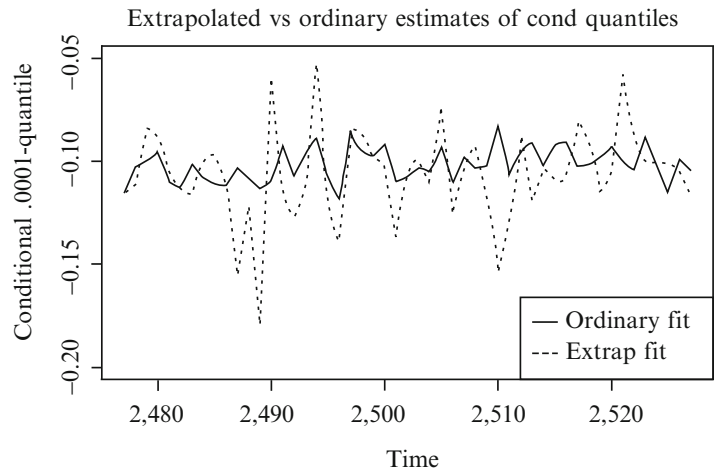
Having characterized the EV index, we can now estimate very extreme quantiles using extrapolation methods. We set the risk level at $\tau = 0.0001$, so that the return falls below $F^{-1}_{Y_t}(0.0001|X_t)$ only once per about 30 years, a very

Extremal Quantiles and Value-at-Risk, Table 1 Estimation results

Coefficients	Estimate	Bias-corrected	90% conf. region
$\tau = 0.001$			
Intercept	−0.08	−0.08	[−0.11, −0.06]
Lag return	0.12	0.01	[−0.19, 0.63]
Oil price	−0.05	−0.05	[−0.42, 0.11]
DJI return	0.71	0.73	[0.05, 1.08]
$\tau = 0.01$			
Intercept	−0.05	−0.05	[−0.05, −0.04]
Lag return	−0.04	−0.06	[−0.16, 0.12]
Oil price	−0.05	−0.06	[−0.17, 0.02]
DJI return	0.49	0.50	[0.24, 0.65]
$\tau = 0.05$			
Intercept	−0.03	−0.03	[−0.03, −0.02]
Lag return	−0.03	−0.04	[−0.10, 0.04]
Oil price	0.01	0.01	[−0.04, 0.06]
DJI return	0.29	0.30	[0.17, 0.38]

Extremal Quantiles and Value-at-Risk, Table 2 Estimation results for the EV index ξ

	Estimate	Bias-corrected	Estimate 90% conf. region
$\tau = 0.005$	0.24	0.22	[0.08, 0.34]
$\tau = 0.01$	0.23	0.17	[0.05, 0.25]
$\tau = 0.025$	0.32	0.24	[0.14, 0.30]
$\tau = 0.05$	0.35	0.23	[0.16, 0.27]



Extremal Quantiles and Value-at-Risk, Fig. 2 Extrapolated and ordinary estimates of the conditional 0.0001-quantile

rare, extreme event. The extrapolated estimates of 0.0001-quantile are obtained using Eq. (41) with $\tau = 0.05$ and $\hat{\xi} = 1/4$. The resulting extrapolation fit $\hat{F}^{-1}_{Y_t}(0.0001|X_t)$ sharply differs from the ordinary fit $X_t\hat{\beta}(\tau)$ obtained by quantile regression. The reason for this is simple: the ordinary fit uses sample data that likely contains no observations on the extreme events defined above. In sharp contrast, the extrapolated fit uses the tail model and a reliably estimated conditional 0.05-quantile to predict the magnitude of such events. The quality of this prediction clearly depends on whether the tail model is accurate (Fig. 2).

Conclusion

This article examines the theory and empirics of extremal quantiles in economics. The theory of extremes provides a set of applicable methods that have generated numerous valuable empirical findings. There is equally promising scope for the use of the extremal conditional quantile methods. The latter methods are new – there are great opportunities for further empirical and theoretical developments.

Acknowledgment We thank Emily Gallagher, Greg Fischer and Raymond Guiteras for their help and valuable comments.

Bibliography

Abrevaya, J. 2001. The effect of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics* 26: 247–259.

Aigner, D.J., and S.F. Chu. 1968. On estimating the industry production function. *American Economic Review* 58: 826–839.

Aigner, D.J., T. Amemiya, and D.J. Poirier. 1976. On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review* 17: 377–396.

Arrow, K., T. Harris, and J. Marschak. 1951. Optimal inventory policy. *Econometrica* 19: 205–272.

Bickel, P., and D. Freedman. 1981. Some asymptotic theory for the bootstrap. *Annals of Statistics* 9: 1196–1217.

Caballero, R., and E. Engel. 1999. Explaining investment dynamics in U.S. manufacturing: A generalized (S,s) approach. *Econometrica* 67: 783–826.

Chernozhukov, V. 2005. Extremal quantile regression. *Annals of Statistics* 33: 806–839.

Chernozhukov, V. 2006. Inference for extremal conditional quantile models, with an application to birthweights. Working paper, Department of Economics, Massachusetts Institute of Technology.

Chernozhukov, V., and L. Umantsev. 2001. Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics* 26: 271–293.

de Haan, L. 1970. *On regular variation and its applications to the weak convergence*, Tract no. 2. Amsterdam: Mathematical Centre.

Dekkers, A., and L. de Haan. 1989. On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics* 17: 1795–1832.

Donald, S.G., and H.J. Paarsch. 2002. Superconsistent estimation and inference in structural econometric models using extreme order statistics. *Journal of Econometrics* 109: 305–340.

Embrechts, P., Klüppelberg, C. and Mikosch, T. 1997. *Modelling extremal events*, vol. 33 of *Applications of mathematics*. Berlin: Springer.

Fama, E.F. 1965. The behavior of stock market prices. *Journal of Business* 38: 34–105.

Fox, M., and H. Rubin. 1964. Admissibility of quantile estimates of a single location parameter. *Annals of Mathematics and Statistics* 35: 1019–1030.

Gnedenko, B. 1943. Sur la distribution limité du terme d' une série alétoire. *Annals of Mathematics* 44: 423–453.

Hill, B.M. 1975. A simple general approach to inference about the tail of a distribution. *Annals of Statistics* 3: 1163–1174.

Jansen, D.W., and C.G. de Vries. 1991. On the frequency of large stock returns: Putting booms and busts into perspective. *Review of Economics and Statistics* 73: 18–24.

Knight, K. 2001. Limiting distributions of linear programming estimators. *Extremes* 4(2): 87–103.

Koenker, R. 2006. *Quantreg: Quantile regression*. R package version 3.90. Online. Available at http://www.r-project.org. Accessed 23 April 2007.

Koenker, R., and G.S. Bassett. 1978. Regression quantiles. *Econometrica* 46: 33–50.

Laplace, P.-S. 1818. *Théorie analytique des probabilités*. Paris: Éditions Jacques Gabay, 1995.

Leadbetter, M.R., G. Lindgren, and H. Rootzén. 1983. *Extremes and related properties of random sequences and processes*. New York/Berlin: Springer.

Longin, F.M. 1996. The asymptotic distribution of extreme stock market returns. *Journal of Business* 69: 383–408.

Mandelbrot, M. 1963. The variation of certain speculative prices. *Journal of Business* 36: 394–419.

Meyer, R.M. 1973. A poisson-type limit theorem for mixing sequences of dependent 'rare' events. *Annals of Probability* 1: 480–483.

Pareto, V. 1964. *Cours d'économie politique*. Genéve: Droz.

Pickands, J. III. 1975. Statistical inference using extreme order statistics. *Annals of Statistics* 3: 119–131.

Portnoy, S., and J. Jurečková 1999. On extreme regression quantiles. *Extremes* 2: 227–243 (2000).
Praetz, V. 1972. The distribution of share price changes. *Journal of Business* 45: 49–55.
Quetelet, A. 1871. *Anthropométrie*. Brussels: Muquardt.
Roy, A.D. 1952. Safety first and the holding of assets. *Econometrica* 20: 431–449.
Sen, A. 1973. *On economic inequality*. New York: Oxford University Press.
Timmer, C.P. 1971. Using a probabilistic frontier production function to measure technical efficiency. *Journal of Political Economy* 79: 776–794.
Zipf, G. 1949. *Human behavior and the principle of last effort*. Cambridge, MA: Addison-Wesley.

Extreme Bounds Analysis

Edward E. Leamer

Abstract

Extreme bounds analysis is a global sensitivity analysis that applies to the choice of variables in a linear regression. Rather than a discrete search over models that include or exclude subsets of the variables, this sensitivity analysis answers the question: how extreme can the estimates be if any linear homogenous restrictions on a selected subset of the coefficients are allowed? When these bounds are too wide to be useful, narrower bounds can be found by restricting the set of prior distributions that underlie the sensitivity analysis.

Keywords

Bayesian econometrics; Extreme bounds analysis; Heteroscedasticity; Nonparametric models and methods; Probability; White-corrected standard errors

JEL Classifications

C13; C14; C21; C41; C51; C53

The analysis of economic data necessarily depends on assumptions that our weak data-sets do not allow us to test. We are forced to choose a limited number of variables in a multivariate analysis, to restrict the functional form, to limit the considered interdependence among observations to special forms, and to make special distributional assumptions. We make these assumptions, not because we believe them, but because we have to. Absent assumptions, our data-sets are utterly useless.

We sometimes put aside the discomfort that our choice of assumptions entails by doing what is conventional, like using a normal distribution, a linear functional form and the same limited set of variables studied by almost everyone else.

We sometimes pretend to treat the problem of choice of assumptions by using 'nonparametric' methods that masquerade as assumption-free. These methods are assumption-free only in Asymptopia, the land where all data-sets are unlimited. To get to the happy land of Asymptopia, we need only let the number of tested assumptions grow in the future more slowly than the number of observations. Then we can be sure to test all assumptions during our journey into the future. But here on Earth we have limited data-sets, and inevitably the way we analyse these data works well for some sets of assumptions and not so well for others. We cannot really know what we are doing unless we can draw some kind of line between the assumptions for which our inferences are valid and the assumptions for which our inferences are not valid. Thus, for example, 'heteroscedasticity-consistent' standard errors are now commonly deployed as if they were corrections for any form of heteroscedasticity. But these corrections of the standard errors, which leave the point estimates unchanged, are an appropriate treatment given the actual limited data only for some forms of heteroscedasticity, not for all. Unfortunately, with these 'nonparametric' methods, the border between the dealt-with assumptions for which the method works and the not-dealt-with assumptions is impossible to draw. Incidentally, these 'heteroscedasticity-consistent' standard errors are often called White-corrected standard errors, to which I respond rhetorically by calling the method 'White-washing'.

If conventions and nonparametric methods are not enough to soften the discomfort with the

assumptions we make, we can always fantasize that the choice of assumptions doesn't really matter. We 'know' the methods we use work well under conditions that are 'close' to the assumptions that underlie them but not well if the departures are great. We hope that the neighbourhood in which the assumptions work is wide enough to encompass the problem at hand. For example, we don't really think the distribution is normal, but how much could that matter for linear regression? Isn't it enough to have symmetric unimodal distributions, shaped 'sort of like' a normal distribution?

Neither conventions, nor nonparametric sleight-of-hand, nor hope that it doesn't really matter form an adequate scientific response to doubt about the assumptions that underlie a data analysis. The correct way to deal with ambiguity in the choice of assumptions that are beyond the range of statistical tests is a sensitivity analysis that demonstrates that our assumptions do not in fact matter much. The most common sensitivity analysis involves the choice of variables in linear regression. Rather than reporting just one regression, many researchers offer a table of results, all based on different subsets of the variables. Typically, all the reported regressions have a common set of 'core' variables but differ depending on whether or not the regressions include selected 'doubtful' variables.

Although it can be comforting to discover that the coefficients of the core variables do not change much when doubtful variables are excluded, this kind of sensitivity analysis leaves open the possibility that there is some combination of doubtful variables that would radically change the result. Has the analyst worked hard enough to find the oddball estimates that these data allow? 'Extreme bounds analysis' answers this question. In an extreme bounds analysis, the computer chooses the linear combinations of doubtful variables that, when included in the regressions along with the core variables, produce the most extreme (minimum and maximum) estimates for the coefficient on a selected core variable. There is no way of fiddling with the doubtful variables that can produce an estimate outside the extreme bounds.

If the extreme bounds interval is small enough to be useful, that is the end of the story, and the result is reported to be 'sturdy'. This would occur, for example, when the core variables and the doubtful variables are 'independent', in which case the coefficients of the core variables don't change at all when doubtful variables are excluded. But quite often with highly correlated economics data these extreme bounds can be uncomfortably wide, and we are forced either to retreat in dismay or to seek some way to make the bounds narrower.

One way of restricting the range of alternative models is to allow only inclusion/exclusion options, not the all linear combinations embodied in the extreme bounds. These inclusion/exclusion restrictions are the basis for the tables of alternative results that are commonly offered as evidence of inferential sturdiness, and the set of alternative estimates thus presented is smaller than the extreme bounds set.

But why? Why restrict to inclusion/exclusion options? Classical inference is not well suited to respond to this 'why?' question since the answer depends on the state of mind of the analyst and since classical inference presumes a researcher and an audience with a 'blank slate'. Moreover, the effect on the inferences of setting a regression coefficient to zero depends on the coordinate system for defining the parameters. (If you use x and z as explanatory variables, and I use $x + z$ and $x - z$, we get different answers.) Indeed, the extreme bounds are formed by setting coefficients to zero in an appropriately defined coordinate system, and therefore restriction to inclusion/exclusion restrictions is a meaningless restriction absent advice on how to define the coordinate system.

A Bayesian analysis can help to choose a coordinate system and can be a basis for a sensitivity analysis with a set of models that is sensibly smaller than the set of models underlying the extreme bounds. Bayesians allow the state of mind to influence the analysis by letting a researcher act as if the vector of regression coefficients on the doubtful variables, $\boldsymbol{\theta}$, comes from a normal distribution with a mean vector $\mathbf{0}$ and covariance matrix $\mathbf{V}_0$, selected by the researcher.

The smaller is the covariance matrix $\mathbf{V}_0$, the more likely are the coefficients $\boldsymbol{\theta}$ to hug close to zero and the more doubtful are the doubtful variables.

To parallel the decision always to include the core variables in the equation, it is natural to deploy a prior probability distribution for the core coefficients with an infinite variance – the blank-slate initial-ignorance option. Then, corresponding to each choice of covariance matrix for the coefficients of the doubtful variables are estimates of the coefficients of the core variables, $\hat{\beta}(\mathbf{V}_0)$. If, for example, the covariance matrix $\mathbf{V}_0$ is set to zero, this is equivalent to assuming the coefficients of the doubtful variables are all zero, and we should be running the regression with all these doubtful variables omitted. Conversely, if the covariance matrix $\mathbf{V}_0$ is set to an 'infinitely large' matrix (unlimited variances), then this is the ignorance option for the doubtful variables, and the way to do the estimation is simply to include all the doubtful variables along with the core variables in the regression.

A 'global' sensitivity analysis in this setting is carried out by building a correspondence between sets of prior covariance matrices for the doubtful variables, V_0, and the corresponding sets of estimates of the coefficients on the core variables, $\hat{\beta}(\mathbf{V}_0)$.

Three possibilities are discussed in Leamer (1978) and Leamer and Chamberlain (1976): (*a*) $\mathbf{V}_0$ unrestricted, (*b*) $\mathbf{V}_0$ diagonal, (*c*) $\mathbf{V}_0$ diagonal with all diagonal elements the same.

1. The extreme bounds apply when the covariance $\mathbf{V}_0$ is any positive semi-definite matrix.
2. The 2^p regressions, found by including different subsets of the p doubtful variables, define the bound when $\mathbf{V}_0$ is a (non-negative) *diagonal* matrix. (Thus the 'right' coordinate system is the one in which the regression coefficients are a priori independent of each other in the sense that knowledge about one doesn't affect your thinking about the others.)
3. A still narrower set of bounds is found by estimating the p + 1 'principal component' regressions with the principal component restrictions ordered by their eigenvalues. This applies when $\mathbf{V}_0$ is proportional to the identity matrix.

Each of these three sets of prior covariance matrices includes the dogmatic priors that set certain linear combinations of the coefficients exactly to zero and also complete ignorance priors that allow certain linear combinations to be completely free. Neither of these two extremes is sensible in practice, since in the first case the data evidence is completely ignored (the restriction is imposed without testing) and in the second case the prior information is completely ignored. These can be excluded to restricting the prior covariance matrix from above and below:

$\mathbf{V}_L < \mathbf{V}_0 < \mathbf{V}_U$ where $\mathbf{A} < \mathbf{B}$ means that the matrix $\mathbf{B} - \mathbf{A}$ is positive definite. The theorem that then applies comes from Leamer (1981, 1982) as is reported below. First, a statement about the posterior mean of the regression coefficient vector.

Theorem (Bayes Estimate) If, conditional on the observable matrix $\mathbf{X}$, and the unobservable parameters β, and σ^2, an observable vector $\mathbf{y}$ is normally distributed with mean $\mathbf{X}\beta$ and covariance matrix $\sigma^2\mathbf{I}$, and if the coefficient vector β comes from a normal distribution with mean $\mathbf{b}_0$ and covariance matrix $\mathbf{V}_0$, then the conditional mean of β given $\mathbf{y}$ is approximately

$$\mathbf{b}_2 = \left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_0^{-1}\right)\left(\mathbf{X}'\mathbf{X}\mathbf{b}/s^2 + \mathbf{V}_0^{-1}\mathbf{b}_0\right)$$

where s^2 is the sample estimate of $\sigma^2 : s^2 = \mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}/(n-k)$, where n is the number of observations and k is the number or regression coefficients and where $\mathbf{b}$ is the ordinary least squares estimator (a solution to the normal Equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}) : \mathbf{b} = (\mathbf{X}'\mathrm{X})^{-1}\mathbf{X}'\mathbf{y}$.

Theorem (Posterior Bounds) Given $\mathbf{V}_L < \mathbf{V}_0 < \mathbf{V}_U$ with $\mathbf{V}_L$ and $\mathbf{V}_U$ positive definite and with $\mathbf{V}_L < \mathbf{V}_0$ signifying that $\mathbf{V}_0 - \mathbf{V}_L$ is positive definite, then the posterior mean $\mathbf{b}_2$ lies in the ellipsoid

$$(\mathbf{b}_2 - \mathbf{f})'\mathbf{H}(\mathbf{b}_2 - \mathbf{f}) < \mathbf{c}$$

where

$$\begin{aligned}\mathbf{H} &= \left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1}\right) \\ &+ \left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1}\right)\left(\mathbf{V}_L^{-1} - \mathbf{V}_U^{-1}\right)^{-1}\left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1}\right)\mathbf{f} \\ &= \left[\left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_L^{-1}\right)\right]^{-1}\left[\mathbf{X}'\mathbf{X}\mathbf{b}/s^2 + \left(\mathbf{V}_L^{-1} - \mathbf{V}_U^{-1}\right)\left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1}\right)^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}/2s^2\right]c \\ &= \left(\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/s^2\right)\left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_U^{-1}\right)^{-1}\left(\mathbf{V}_L^{-1} - \mathbf{V}_U^{-1}\right)\left(\mathbf{X}'\mathbf{X}/s^2 + \mathbf{V}_L^{-1}\right)^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}/4s^2\end{aligned}$$

See Also

- ▶ Bayesian Econometrics
- ▶ Data Mining

Bibliography

Leamer, E. 1978. *Specification searches: Ad Hoc Inference with Non experimental data*. New York: John Wiley and Sons.

Leamer, E. 1981. Sets of estimates of location. *Econometrica* 49: 193–204.

Leamer, E. 1982. Sets of posterior means with bounded variance priors. *Econometrica* 50: 725–736.

Leamer, E., and G. Chamberlain. 1976. Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society, Series B* 38: 73–84.

Extreme Poverty

Jeffrey D. Sachs and Gordon C. McCord

Abstract

Households living in extreme poverty face deprivations that cost millions of lives annually. Ending extreme poverty requires an understanding of poverty traps, including the effects of adverse biophysical and geographical factors, a lack of resources required for the investments needed to escape poverty, and poor governance. Policies must focus both on promoting market-oriented economic growth and on directly addressing the needs of the poor. Foreign aid will be required to finance interventions that poor countries cannot finance themselves, and aid to well-governed poor countries should be increased, consistent with the rich-country promise of 0.7% of GNP as official development assistance.

Keywords

Business capital; Capabilities; Corruption; Democracy; Development assistance; Diffusion of technology; Economic growth; Education; Extreme poverty; Family planning; Fertility; Food security; Foreign aid; Geography and economic development; Governance; Health care; Human capital; Human Development Index; Human rights approach to poverty alleviation; Infant mortality; Innovation; Knowledge capital; Millennium development goals; Moderate poverty; Natural capital; Nutrition; Patents; Poverty; Poverty alleviation; Poverty lines; Poverty traps; Public goods; Public infrastructure; Public institutional capital; Relative poverty; Research and development; Sen, A; State failure; Sub-Saharan Africa; Subsistence; Washington Consensus; Water stress

JEL Classification

C11; C14

There are many definitions of poverty, as well as intense debates about the exact numbers of the poor, where they live and how their numbers are changing over time. As a matter of definition, it is useful to distinguish between three degrees of poverty. Extreme (or absolute) poverty, moderate

poverty and relative poverty. Extreme poverty can be thought of as 'poverty that kills', meaning that households cannot reliably meet basic needs for survival. Households living in extreme poverty are chronically undernourished, unable to access health care, lacking the amenities of safe drinking water and sanitation, unable to afford education for some or all of the children, and perhaps lacking rudimentary shelter – a roof to keep the rain out of the hut, a chimney to remove the smoke from the cooking stove – and basic articles of clothing such as shoes. Such deprivations cost lives, by the millions, every year. Life expectancy is considerably lower and mortality rates are considerably higher in countries in which large proportions of the population live in extreme poverty.

Unlike moderate and relative poverty, extreme poverty currently occurs only in developing countries. Moderate poverty generally refers to the conditions of life in which basic needs are met, but only barely. Relative poverty is generally construed as a household income level below a given proportion of average national income. The relatively poor, in high-income countries, lack access to cultural goods, entertainment, recreation, and to quality health care, education, and other perquisites of social mobility. They may also live outside of the 'mainstream' of social life, and thus without dignity and social respect.

In order to estimate the number of extreme poor, most analysts use a poverty line – a level of income below which the person is 'extremely poor' by some definition. Most countries set their own poverty lines, based on the per capita cost of a consumption basket that attempts to measure basic needs. Since the poorest people in poor countries spend most of their money on food, most of the basket used for national poverty lines consists of food, usually in terms of meeting a minimum intake of 2000 calories (Deaton 2004). These poverty lines are surely imperfect: they suffer from the measurement error inherent in household surveys; they are rarely updated with regards to spending on nutrition; they do not account for differences in rural versus urban calorie consumption; and they do not capture all dimensions of extreme poverty (for example, access to health care, safe water, sanitation, education or political voice).

Moreover, they can lead to undesirable policy results (a person just below the poverty lines could be treated very differently from someone just above the line, despite having almost equal incomes). Governments judged solely according to the number of people below the poverty line could choose to focus only on those closest to the line and ignore the poorest of the poor. Finally, as Nobel Laureate Amartya Sen has emphasized, poverty should be defined more broadly than having a low income; rather, it is the absence of basic capabilities to function in society. This could include not only income poverty (involving a lack of food, clothing, or shelter), but also lack of access to public goods, social standing, and political participation. Despite these shortcomings, most dimensions of extreme poverty that people would like to improve are correlated with household income, thus making a poverty line a helpful, though rough, first approximation of poverty rates. Measures that combined household income with provisions of public goods (disease control, public health, primary education) would surely be preferable.

In the late 1980s, and especially with the 1990 *World Development Report,* the World Bank introduced a single measure of extreme poverty – an income of one dollar per day or less (in 1985 purchasing power parity, PPP, dollars) – in order to compare rates of extreme poverty across countries and to track extreme poverty over time. The one dollar per day number was chosen since it corresponds roughly to the highest national poverty rate among low-income countries (around 360 dollars per year). In 2000, the World Bank used improved PPP estimates to adjust its global poverty line to 1.08 dollars per person per day (in 1993 PPP dollars). This global extreme poverty line has been criticized by some for not being high enough and thus undervaluing the needs of the poor (Pritchett 2003) and by others for being too arbitrary and detached from the country-specific needs of the poor (Srinivasan 2004). Nevertheless, it provides a useful, albeit highly imperfect, measuring tool to look at extreme poverty around the world.

Extreme Poverty, Table 1 Number of poor people by region, 1981–2001

	$ 1.08 per day (million) 1987							
	1981	1984	1990		1993	1996	1999	2001
East Asia	795.6	562.2	425.6	472.2	415.4	286.7	281.7	271.3
Of which China	633.7	425.0	308.4	374.8	334.2	211.6	222.8	211.6
Eastern Europe and Central Asia	3.1	2.4	1.7	2.3	17.4	19.8	29.8	17.6
Latin America and Caribbean	35.6	46.0	45.1	49.3	52.0	52.2	53.6	49.8
Middle East and North Africa	9.1	7.6	6.9	5.5	4.0	5.5	7.7	7.1
South Asia	474.8	460.3	473.3	462.3	476.2	461.3	428.5	431.1
Of which India	382.4	373.5	369.8	357.4	380.0	399.5	352.4	358.6
Sub-Saharan Africa	163.6	198.3	218.6	226.8	242.3	271.4	294.0	315.8
Total	1481.8	1276.8	1171.2	1218.5	1207.5	1096.9	1095.1	1092.7

Source: Chen and Ravallion (2004, p. 153)

Another important indicator for poverty is the Human Development Index (HDI), published by the United Nations Development Programme (UNDP) since 1990. The UNDP sought to incorporate the multidimensional aspects of poverty into a new indicator, and to emphasize that development should expand human capabilities, particularly those that are universally valued and basic to life: the capability to lead a long and healthy life, to be knowledgeable, and to have access to the resources needed for a decent standard of living (UNDP 2004). The result was the HDI, which averages normalized 0–1 indexes for income per capita, life expectancy, and education school enrolment and literacy. Countries classified as 'low human development' have a very strong overlap with those countries that have a high proportion of the population living under one dollar per day according to the World Bank.

Where Are the Poor?

The most recent estimates of extreme poverty around the world (using the one dollar per day estimate) were made by Shaohua Chen and Martin Ravallion at the World Bank (see Table 1). They estimated that roughly 1.1 billion people were living in extreme poverty in 2001, down from 1.5 billion in 1981 (Chen and Ravallion 2004). The overwhelming share of the world's extreme poor, 93% in 2001, live in three regions, East Asia, South Asia and Sub-Saharan Africa. Since 1981, the absolute numbers of extreme poor have risen in Sub-Saharan Africa, but have fallen in East Asia and South Asia. In terms of proportions, nearly half Africa's population is judged to live in extreme poverty, and that proportion has risen slightly over the period. The proportion of the extreme poor in East Asia has plummeted, from 58% in 1981 to 15% in 2001; in South Asia the progress has also been marked, although slightly less dramatically, from 52 to 31%. Latin America's extreme poverty rate is around 10%, and relatively unchanged; Eastern Europe's rose from a negligible level in 1981 to around 4% in 2001, the results of the upheavals of Communist collapse and economic transition to a market economy. It is worth noting that these numbers are debated heatedly; other researchers have relied on national income accounts, which tend to show somewhat faster progress in the reduction of Asian poverty, and sometimes very different estimates for the total amount of people living in extreme poverty (Sala-i-Martin 2002; Bhalla 2002). The general picture, however, remains true in all these studies: extreme poverty is concentrated in East Asia, South Asia and Sub-Saharan Africa. It is rising in Africa in absolute numbers and as a share of the population, while it is falling both in absolute numbers and as a proportion of the population in the Asian regions.

There are some defining circumstances specific to the poorest of the poor. They are found mainly in rural areas (though with a growing proportion in the cities); the rural poor tend to have fewer opportunities to earn income, have less access to education and health care, and are often more vulnerable to the forces of nature. The extreme

poor face challenges almost unknown in the rich world today – malaria, famines, lack of roads and motor vehicles, great distances to regional and world markets, lack of electricity and modern cooking fuels. Women tend to be at a disadvantage compared with men, since they often have less access to property rights (land ownership, inheritance), and since they bear the physical burden of lack of infrastructure (collecting water and fuel wood at great distances). Girls have historically received less primary and secondary education than boys. Labour markets often discriminate against women, and women tend to work longer when one counts unpaid labour at home. Domestic violence continues to burden the lives of millions of women around the world (World Bank 2001). Finally, large pockets of poverty exist within many countries due to racial and ethnic discrimination, or low social (for example, caste) status.

Consequences of Extreme Poverty

When individuals suffer from extreme poverty and lack the meagre income needed to cover even basic needs, a single episode of disease, a drought, or a pest that destroys a harvest can be the difference between life and death. In households suffering from extreme poverty, life expectancy is often around half that in the high- income world, 40 years instead of 80 years. It is common that, in the poorest countries of Sub-Saharan Africa, of every 1000 children born more than 100 die before their fifth birthday, compared with fewer than ten in the high-income world. An infant born in Sub-Saharan Africa today has only a one-in-three chance of surviving to age 65.

At the most basic level, the poorest of the poor lack the minimum amount of capital necessary to get a foothold on the first rung of the ladder of economic development. The extreme poor tend to lack six major kinds of capital:

- Human capital: health, nutrition, and skills – education – needed for each person to be economically productive.
- Business capital: the machinery, facilities, motorized transport used in agriculture, industry and services.
- Infrastructure: roads, power, water and sanitation, airport and seaports, and telecommunications systems, which are critical inputs into business productivity.
- Natural capital: arable land, healthy soils, biodiversity, and well-functioning ecosystems that provide the environmental services needed by human society.
- Public institutional capital: commercial law, judicial systems, government services and policing that underpin the peaceful and prosperous division of labour.
- Knowledge capital: the scientific and technological know-how that raises productivity in business output and the promotion of physical and natural capital.

Importantly, the poorest of the poor tend to have higher fertility rates, for several reasons. Infant mortality rates are high when there are inadequate health services, so high fertility provides 'insurance' to parents that they will succeed in raising a child who will survive to adulthood. In rural areas, children are often perceived as economic assets who provide supplementary labour for the farm household. Poor and illiterate women have few job opportunities away from the farm, and so may place a low value on the opportunity (time) costs of bringing up children. In addition, women are frequently unaware of their reproductive rights (including the right to plan their families) and lack access to reproductive health information, services, and facilities, leading to high unmet demands for contraception in low-income countries and among poorer members of all developing countries. Finally, poor households lack the income to purchase contraceptives and family planning, even when they are available. For these reasons, high fertility rates are prevalent among families living in extreme poverty, resulting in very low investments in the health and education of each child (what is known as the quantity–quality trade-off).

Poor and hungry societies are much more likely than high-income societies to fall into

violent conflicts over scarce vital resources, such as watering holes and arable land – and over scarce natural resources, such as oil, diamonds and timber (United Nations, 2004). This relationship between violence and high rates of extreme poverty holds with a high degree of statistical significance. A country with a civil war within its borders typically has only one-third of the per capita income of a country with similar characteristics but at peace. Moreover, poor countries – even those not in conflict – risk conflict in the future. A country with a per capita income of 500 dollars is about twice as likely to have a major conflict within 5 years as a country with an income of about 4000 dollars per capita (UN Millennium Project 2005). In addition, low economic growth rates are associated with higher risks of new conflict; one study finds that a negative growth shock of 5% increases the risk of civil war by 50% in the following year, and that economic conditions are probably the most important determinants of civil conflict in Sub- Saharan Africa (Miguel et al. 2004). The most comprehensive study of state failure, carried out by the State Failure Task Force established by the Central Intelligence Agency in 1994, confirms the importance of the economic roots of state failure (defined as revolutionary war, ethnic war, genocide, politicide, or adverse or disruptive regime change). The Task Force studied all 113 cases of state failure between 1957 and 1994 in countries of half a million people or more, and found that the most significant variables explaining these conflicts were the infant mortality rate (suggesting that overall low levels of material well-being are a significant contributor to state failure), openness of the economy (more economic linkages with the rest of the world diminish the chances of state failure), and democracy (democratic countries show less propensity to state failure than authoritarian regimes). The linkage to democracy also has a strong economic dimension, however, because research has shown repeatedly that the probability of a country's being democratic rises significantly with its per capita income level. In refinements of the basic study, the Task Force found that in Sub-Saharan Africa, where many societies live on the edge of subsistence, temporary economic setbacks (measured as a decline in gross domestic product per capita) were significant predictors of state failure (State Failure Task Force 1999). Similar conclusions have been reached in studies on African conflict, which find that poverty and slow economic growth raise the probability of conflict.

Theories of Extreme Poverty

For decades, observers have tried to explain why extreme poverty persists. Many theories have looked for single-factor explanations for a lack of economic growth, often grounded in racist beliefs (poor countries do not grow because their cultures, races, or religions fail to promote economic growth). The increasing number of success stories of growth proved all these theories to be wrong. However, despite the complexity of an economy and the number of things that can go wrong, single-factor explanations persist. The most common is that poverty is a result of corrupt leadership, which impedes modern development.

Governance is indeed important: economic development stalls when governments do not uphold the rule of law, pursue sound economic policy, make appropriate public investments, manage a public administration, protect basic human rights, and support civil society organizations – including those representing poor people – in national decision-making. Importantly, long-term poverty reduction in developing countries will not happen without sustained economic growth, which requires a vibrant private sector. Government, therefore, needs to provide the economic policy framework and the support that the private sector needs to grow.

However, many well-governed poor countries may be too poor to help themselves out of extreme poverty. Many well-intentioned governments lack the fiscal resources to invest in infrastructure, social services, environmental management, and even the public administration necessary to improve governance. Further, dozens of heavily indebted poor and middle-income countries have been forced by creditor governments to spend large proportions of their limited tax receipts on

debt service, undermining their ability to finance vital investments in human capital and infrastructure. The reason these poor countries cannot grow is not poor governance, but a poverty trap. They lack the basic infrastructure, human capital, and public administration – the foundations for economic development and private sector-led growth. Without roads, soil nutrients, electricity, safe cooking fuels, clinics, schools, and adequate and affordable shelter, people are chronically hungry, burdened by disease and unable to save. As mentioned above, fertility rates tend to be high, preventing families from investing enough in each child. Without adequate public sector salaries and information technologies, public management is chronically weak. For all of these interlocking reasons, these countries are then unable to attract private investment flows or retain their skilled workers, and can therefore find themselves with low or negative growth. In short, they are stuck in a poverty trap.

The concept of a low-level poverty trap is a long-standing hypothesis in the theories of economic growth and development. The earliest mathematical formalization was by Nelson (1956), who put emphasis on demography. The theoretical possibility of poverty traps in the neoclassical growth model is covered briefly in the economic growth textbook by Barro and Sala-i-Martin (1998), which also discusses briefly the possible case for large-scale development assistance to overcome such traps. The connection of a low-level trap to subsistence consumption needs is spelled out in Ben-David (1998), and connections to agriculture and education are described in the *World Economic and Social Survey* (UN 2000). Two recent empirical studies claiming that such poverty traps exist in poor countries are UNCTAD (2002) and Bloom et al. (2003). A close look at a poverty trap in Sub-Saharan Africa is in Sachs et al. (2004).

An often overlooked characteristic of poverty is that some countries and regions are clearly more vulnerable than others to falling into a poverty trap. While a history of violence of colonial rule or poor governance can leave any country bereft of basic infrastructure and human capital, physical geography plays special havoc with certain regions. Some regions need more basic infrastructure than others simply to compensate for a difficult physical environment. Some of the barriers that must be offset by investments include adverse transport conditions (landlocked economies, small island economies far from major markets, inland populations far from coasts and navigable rivers, populations living in mountains, long distances from major world markets, very low population densities); adverse agro-climatic conditions (low and highly variable rainfall, lack of suitable conditions for irrigation, nutrient-poor and nutrient-depleted soils, vulnerability to pests and other post-harvest losses, susceptibility to the effects of climate change); adverse health conditions (high ecological vulnerability to malaria and other tropical diseases, high AIDS prevalence); and other adverse conditions (lack of domestic energy sources, small internal market and lack of regional integration, vulnerability to natural hazards, artificial borders that cut across cultural and ethnic groups, proximity to countries in conflict). Adam Smith was acutely aware of the role of geography in hindering economic development. He stressed, in particular, the advantages of proximity to low-cost, sea-based trade as critical, noting that remote economies would be the last regions to achieve economic development. More recent studies have found statistical significance of these relationships between geography and economic outcome (Gallup et al. 1999a, b; Mellinger et al. 2000; Sachs and Gallup 2001).

In the rich countries of North America, Western Europe and East Asia, the process of massive investment in research and development, leading to sales of patent-protected products to a large market, stands at the core of economic growth. Advanced countries are typically investing 2% or more of their gross national product directly into the research and development process, and sometimes more than 3%. That investment is very sizeable, with hundreds of billions of dollars invested each year in research and development activities. Moreover, these investments are not simply left to the market. Governments invest heavily, especially in the early stages of R&D. In most poor countries, especially smaller ones, the innovation process usually never gets started.

Potential inventors do not invent because they know that they will not be able to recoup the large, fixed costs of developing a new product. Impoverished governments cannot afford to back the basic sciences in government laboratories and in universities. The result is an inequality of innovative activity that magnifies the inequality of global incomes. While the innovation gap is reduced in the case of some poor countries through technological diffusion, even diffusion is limited in the poorest countries, because they face distinctive ecological problems not addressed by 'rich-world science' (for example, tropical diseases and tropical farming systems), because they cannot afford high-tech capital goods and because they fail to attract foreign businesses that would bring the technology with them.

Policy Responses

Theories on how to tackle extreme poverty are varied and controversial. For the most part, they can be divided into two camps: strategies that focus on promoting market- oriented economic growth, and strategies that focus on directly addressing the needs of the poor. Of course the two approaches can be combined. The Washington Consensus, a set of policy recommendations especially prevalent from 1980 to the late 1990s, embodies the first type, with its focus on macroeconomic stability, greater economic openness to trade and investment, and improved environment for private business. The idea was that these policies would lead growth of the private sector, thus increasing demand for labour and thereby improving the welfare of the poor.

A second set of strategies focuses instead on providing what the poor need in order to increase their productivity. These investments in 'human development' argued for directing health and education investments towards the poor, and providing social safety nets. Many of these strategies became popular in the 1990s as a reaction to the Washington Consensus. There were three kinds of critiques. One held that growth would not be achieved with market reforms alone, because of the poverty trap. A second held that growth must in any event be combined with increased public investments, for example for health and education. A third, and more extreme position, held that growth per se would have adverse effects on the poorest of the poor. For example, the 1996 *Human Development Report* warned that in some cases growth can fail to create jobs and provide benefits, and can even increase empowerment of the rich, wreck cultural identities and destroy the environment (UNDP, 1996).

Numerous studies have shown that growth *does*, in fact, tend to be good for the poor (Dollar and Kraay 2002; Roemer and Gugerty 1997; Gallup et al. 1999a, b). Yet growth may not be achievable for countries trapped in poverty, and growth may not be sufficient to enable the poorest of the poor to meet their basic needs. The emerging consensus is that *both* economic growth and direct investments for the poor are necessary, in order to break the poverty trap and to provide vital public goods. International institutions are paying more attention than before to the possibility of poverty traps, and to the non-income dimensions of extreme poverty (for example, health and education). Five of the eight Millennium Development Goals (the world's time-bound and quantified targets for addressing extreme poverty, discussed below) are about promoting health and education, and individual countries are giving more priority to these broader measures than ever before.

Another dimension of the fight against extreme poverty is referred to as the rights-based approach. The guarantee that all people can live in dignity and meet their basic needs is also a basic human right – the right of each person on the planet to health, education, shelter and security as pledged in the Universal Declaration of Human Rights and various UN covenants, treaties and intergovernmental documents (such as the UN Millennium Declaration). The human rights approach seeks to use national and international human rights accountability mechanisms to monitor action on behalf of a human right rather than a development target. Economic evaluations often measure whether a given policy action contributes to reaching a target. Conceived in terms of rights, the same evaluation would measure not only those

reached by a given action, but several other considerations as well: (*a*) the numbers not being reached; (*b*) the empowerment of the poor to achieve their rights; (*c*) the protection of these rights in legislation; and so forth. To date, there has been insufficient effort to integrate development planning with a human rights framework, even though such integration has tremendous potential and relevance.

Since the creation of the United Nations in 1945, the international system has been working to reduce poverty around the world, but often with results that fall short of laudable rhetoric. In January 1961, the United Nations resolved that the decade of the 1960s would be the Decade of Development. US President Kennedy launched the decade at the UN in New York. Earlier, in his inaugural address as President, he had signalled a new sense of purpose in international affairs. He declared: 'To those peoples in the huts and villages of half the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves' (History Place 2007). The second Development Decade resolved to emphasize measures deliberately targeted at the poor – to help them meet their basic needs for food, water, housing, health and education. The UN held a series of international conferences: on environment (Stockholm, 1972); population (Bucharest, 1974); food (Rome, 1974); women (Mexico City, 1975); human settlements (Vancouver, 1976); employment (Geneva, 1976); water (Mar del Plata, 1977); and desertification (Nairobi, 1977). In 1978, the governments of the world came together to sign the Alma Ata Declaration that promised 'Health for All by 2000', a promise the world failed miserably in delivering. The 1980s – the third Development Decade – were very difficult for developing countries as they suffered from a worldwide recession that hit the developing world and debtor countries with special force. Nevertheless, important improvements were made in some areas, such as nutrition, access to safe drinking water, and reductions in child mortality. One result was the international conference held in 1990 under the auspices of UNDP, UNESCO, UNICEF and the World Bank in Jomtien (Thailand), which set the target of 'Education for All by the Year 2000', another goal not met.

The 1990s also became a decade in which the response of the UN system to the flagging development movement was to embark on a series of global conferences. The UN Conference on Environment and Development (Rio de Janeiro, 1992) was followed by conferences on nutrition (Rome, 1992); human rights (Vienna, 1993); population and development (Cairo, 1994); social development (Copenhagen, 1995); women (Beijing, 1995), human settlements (Istanbul, 1996). The decade ended with the landmark Millennium Summit in 2000, which resulted in the Millennium Development Goals (MDGs), and the Financing for Development Conference in Monterrey in 2002, where rich countries renewed their pledge to provide 0.7% of their GDP in foreign aid. Also relevant was the Brussels Programme of Action for the Least Developed Countries, which suggests that they require greatly increased official development assistance, since private capital flows will not finance needed public investments. The programme outlines several priority areas for cooperation including human and institutional resource development, removing supply side constraints and enhancing productive capacity, protecting the environment, and attaining food security and reducing malnutrition.

As the UN Millennium Project has pointed out, the Millennium Development Goals are the most broadly supported, comprehensive, and specific poverty reduction targets the world has ever established, so their importance is manifold. For the international political system, they are the fulcrum on which development policy is currently based. For the billion-plus people living in extreme poverty, they represent the means to a productive life (UN Millennium Project 2005). Besides aiming to reduce the 1990 proportion of people in extreme poverty by half by 2015, the MDGs tackle poverty in its many dimensions – income poverty, hunger, disease, lack of adequate shelter, and exclusion – while promoting gender equality, education and environmental sustainability. Thus, while supporting the need for economic growth, the MDGs emphasize that the growth needs to be pro-poor. In 2005, the UN

Millennium Project presented the Secretary General with 'A Practical Plan to Achieve the Millennium Development Goals', which outlined specific interventions to address the multiple causes of poverty traps in poor countries around the world (UN Millennium Project 2005). Moreover, it emphasized that foreign aid will be needed to finance the interventions that the poor countries cannot finance themselves. In the case of well-governed poor countries, the report recommended that foreign assistance should be scaled up immediately, significantly, and on a sustained basis, consistent with the promise of 0.7% of GNP as official development assistance.

Prospects

There are reasons to be optimistic about the elimination of extreme poverty on the planet. Economic development has lifted more than 100 million people out of extreme poverty since the mid-1990s, and the pace is probably accelerating in Asia. While the population of developing countries rose from about 4 billion people to about 5 billion, average per capita incomes rose by more than 21%. With 130 million fewer people in extreme poverty in 2001 than a decade before, the proportion of people living on less than 1 dollar a day declined by 7% points, from 28 to 21%.

Despite the good news, however, Africa remains mired in seemingly intractable extreme poverty. Africa faces difficult structural challenges (very high transport costs and small markets, low-productivity agriculture, very high disease burden, a history of adverse geopolitics, and slow diffusion of technology from abroad), but, in countries where governments are committed, these challenges can be overcome if addressed through an intensive programme that directly confronts them (Sachs et al. 2004). Ending the poverty trap in Africa and meeting the MDGs will require a comprehensive strategy for public investment in conjunction with improved governance. The good news is that the amount of investment required, although out of reach of African governments alone, is within the amount already promised in foreign aid by the rich countries (UN Millennium Project 2005).

One final point is that a sustained reduction in extreme poverty requires tackling long-term challenges that the human family faces, in particular environmental challenges. Raising the incomes of billions of people around the world is surely desirable. Nevertheless, the increased income will come with increased demand for food, energy, and consumer goods, which may push our planet's already stressed ecosystems beyond what they can support. As the world works towards eliminating extreme poverty, it must do so with a conscious plan to limit the environmental burden that humanity places on the planet. Moreover, in many cases, the environmental challenges (such as water stress) may prove to be the biggest barriers to poverty reduction even in the short term.

See Also

- ▶ Foreign Aid
- ▶ Poverty
- ▶ Poverty Alleviation Programmes
- ▶ Poverty Lines
- ▶ Poverty Traps
- ▶ Sen, Amartya (Born 1933)

Bibliography

Barro, R., and X. Sala-i-Martin. 1998. *Economic growth*. Cambridge, MA: MIT Press.

Ben-David, D. 1998. Convergence clubs and subsistence economies. *Journal of Development Economics* 55: 155–171.

Bhalla, S. 2002. *Imagine there's no country – Poverty, inequality and growth in the era of globalization*. Washington, DC: Institute for International Economics.

Bloom, D., D. Canning, and J. Sevilla. 2003. Geography and poverty traps. *Journal of Economic Growth* 8: 355–378.

Chen, S., and M. Ravallion. 2004. How have the world's poorest fared since the early 1980s? *World Bank Research Observer* 19: 141–170.

Deaton, A. 2004. Measuring poverty. Working paper, Research program in development studies. Princeton University.

Dollar, D., and A. Kraay. 2002. Growth is good for the poor. *Journal of Economic Growth* 7: 195–225.

Gallup, J., S. Radelet, and Warner, A. 1999. *Economic growth and the income of the poor*. CAER II discussion paper No. 36. Harvard Institute for International Development.

Gallup, J.L., J.D. Sachs, and A.D. Mellinger. 1999b. Geography and economic development. *International Regional Science Review* 22: 179–232.

History Place. 2007. *Great speeches collection: John F. Kennedy inaugural address*. Online. Available at http://www.historyplace.com/speeches/jfk-inaug.htm. Accessed 25 Apr 2007.

Mellinger, A., J.D. Sachs, and J. Gallup. 2000. Climate, coastal proximity, and development. In *Oxford handbook of economic geography*, ed. G.L. Clark, M.P. Feldman, and M.S. Gertler. New York: Oxford University Press.

Miguel, E., S. Satyanath, and E. Sergenti. 2004. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* 112: 725–753.

Nelson, R. 1956. A theory of the low-level equilibrium trap in underdeveloped economies. *American Economic Review* 46: 894–908.

Perkins, D.H., S. Radelet, and D.L. Lindauer. 2006. *Economics of development*, 6th ed. New York: W.W. Norton.

Pritchett, L. 2003. *Who is not poor? Proposing a higher international standard for poverty*. Working paper No. 33. Center for Global Development.

Roemer, M., and Gugerty, M.K. 1997. *Does economic growth reduce poverty?*. CAER II discussion paper No. 5. Harvard Institute for International Development.

Sachs, J.D. 2005. *The end of poverty*. New York: Penguin.

Sachs, J.D., and J.L. Gallup. 2001. The economic burden of malaria. *American Journal of Tropical Medicine and Hygiene* 64(supplement): 85–96.

Sachs, J.D., J.W. McArthur, G. Schmidt-Traub, M. Kruk, C. Bahadur, M. Faye, and G. McCord. 2004. Ending Africa's poverty trap. *Brookings Papers on Economic Activity* 2004(1): 117–240.

Sala-i-Martin, X. 2002. *The world distribution of income*, Working paper, vol. 8933. Cambridge, MA: NBER.

Srinivasan, T.N. 2004. The unsatisfactory state of global poverty estimation. In *In focus: Dollar a day: How much does it say?* Brasilia: International Poverty Centre, UNDP.

State Failure Task Force. 1999. *Environmental change and security project report*, vol. 5. Washington, DC: Woodrow Wilson Center.

UN (United Nations). 2000. *World economic and social survey 2000*. New York: United Nations.

UN. 2004. *A more secure world: Our shared responsibility. Report of the secretary-general's high-level panel on threats, challenges and change*. New York: United Nations.

UN Millennium Project. 2005. *Investing in development: A practical plan to achieve the Millennium Development Goals*. New York: Earthscan.

UNCTAD (United Nations Conference on Trade and Development). 2002. *The least developed countries report 2002: Escaping the poverty trap*. Geneva: UNCTAD.

UNDP (United Nations Development Programme). 1996. *Human development report 1996*. New York: UNDP/Oxford University Press.

UNDP. 2004. *Human development report 2004: Cultural liberty in today's diverse world*. New York: UNDP/Oxford University Press.

World Bank. 1990. *World development report 1990: Poverty*. Washington, DC: World Bank.

World Bank. 2001. *Engendering development*. New York: Oxford University Press.