L

Laboratory Financial Markets

Daniel Friedman

Abstract

Small-scale financial markets have been studied in the laboratory for more than two decades. Typically, 6–20 human subjects buy and sell units of a single asset whose dividends extend over several periods and/or are uncertain. Such markets permit direct observation of informational efficiency, and allow sharp tests of theoretical predictions. They also provide test beds for policy initiatives, new market formats and automated trading strategies.

Keywords

Arbitrage; Asset price formation; Behavioural finance; Bid–ask markets; Bubbles; Call markets; Continuous double auction; Efficient markets hypothesis; Equity premium; Experimental economics; Financial market anomalies; Information aggregation; Iowa Electronic Market; Laboratory financial markets; Learning; Market institutions; Posted offer; Prediction markets; Prospect theory; South Sea bubble; Stationary repetition; Tulipmania

JEL Classifications

C9

Laboratory financial markets allow human subjects to trade assets under conditions controlled by the researcher. By varying the conditions – such as the trading format, or the timing and content of private information – the researcher can make direct and sharp inferences.

Such inferences are crucial to achieve insight into the ongoing debate about the importance of behavioural anomalies in financial markets (see section "► Behavioural Finance"). Efficient markets and related theories provide a satisfying explanation for many of the properties of modern financial markets, but they are hard to reconcile with well documented 'market anomalies' such as home bias, the large equity premium and excessive volatility. Should financial economists force a reconciliation, or should they embrace prospect theory and other behavioural theories?

These issues are not just academic. Since the collapse of the Soviet bloc around 1990, a dominant share of the world economy has relied on financial markets to choose its economic future. If the efficient markets theory is wrong, and asset prices do not necessarily reflect all available information, then major restructuring may be in order. Perhaps the global economy would be stronger with information disclosures that cater to our behavioural idiosyncrasies, or even with non-market allocation of investment.

Laboratory asset markets inform the debate by offering evidence that complements field data. The strength of experimental methodology is that the researcher can precisely control

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

information, public and private, and can elicit beliefs as well as track offers, transactions and allocations. Thus, in a simplified setting, researchers can systematically dissect the process of asset price formation. In conjunction with theory and field empirical work, laboratory investigations help us understand how financial markets really work.

Early Laboratory Markets

Experimental economics cut its teeth on laboratory commodity markets. Reacting to Edward Chamberlin's casual classroom experiments, Vernon Smith pioneered the scientific study of markets in the laboratory. He refined the idea of *induced value and cost*: the experimenter promises to pay a subject the amount v if she buys a unit, and charges another subject the amount c if he sells a unit. If they transact at price p, she earns $v - p$ and he earns $p - c$, generating surplus of $v - c$. The payments are in cash and large enough for the subjects to take seriously.

Smith introduced *stationary repetition* – several consecutive trading periods with the same endowed values and costs but no carry-over from one period to the next, so that subjects have the opportunity to adapt to the trading environment. He also brought the *continuous double auction* (CDA) market (sometimes referred to as the double oral auction) format into the laboratory: traders can make public, committed offers to buy and to sell and can accept others' offers at any time during a trading period. Variants of the CDA format predominate in modern financial markets, including the New York Stock Exchange (NYSE), NASDAQ, and the Chicago Mercantile Exchange.

Numerous laboratory studies, beginning with Smith (1962), show that CDA markets with only a few buyers and sellers (say, four of each) reliably produce highly efficient outcomes, where efficiency is defined as the fraction of potential surplus in the market that is captured by the buyers and sellers. Typically, over 95 per cent of total surplus is realized after a few periods of stationary repetition.

Such perishable commodity markets provide no interesting role for time or uncertainty, both important dimensions of financial assets. Laboratory financial markets should allow two-way traders who can both buy and sell, and who trade assets with a payout that is uncertain and/or carries over several periods. Experimenters at Caltech first introduced such markets in the early 1980s. For example, Plott and Sunder (1982) created a single period asset that was traded by six uninformed traders, who knew only that one of two states would occur with given probabilities independently each period, and six informed traders, who knew the realized state. Both informed and uninformed traders were distributed evenly across three types of state-contingent dividend schedules. Within a few periods, prices became highly efficient, and the trading patterns demonstrated that the market fully disseminated the private information. About the same time, several teams of researchers found very efficient asset prices in laboratory markets with assets paying individual- and state-contingent dividends over several trading periods. These and other early laboratory experiments demonstrated that futures and options contracts can speed convergence towards efficient asset prices. See Sunder (1995) for a thorough survey.

The main lesson from these studies is that financial markets can process information very efficiently. As Hayek (1945) conjectured, markets can fully aggregate and disseminate dispersed private information, and can do so quite rapidly. A few bids and asks in the CDA suffice to fully inform experienced traders, dealing appropriate assets, in moderately complex environments.

Dissecting Financial Markets

These positive early results encourage us to look more deeply at how financial markets process information. The process has several logical stages. Investors and other participants acquire relevant information from diverse sources, public and private. Individual investors incorporate the information into their beliefs about future asset

prices. Acting on their beliefs, investors try to buy assets they expect to appreciate relatively rapidly and to sell assets that they expect to do less well. Their buy and sell orders in turn produce observable market outcomes such as asset price and trading volume. The market outcomes provide further public information for investors, other new information arrives from time to time, and so the process continues. We now know that the process can work quite well in favourable circumstances. But even the early laboratory studies show that it is sometimes fallible. When and where might it go wrong?

Each stage of the process can be examined in the laboratory and compared with theoretical predictions. Cognitive scientists focus on the first stage, the formation of beliefs given arriving information, and have documented many biases that might distort beliefs. Examples include overconfidence, the gambler's fallacy (believing that a coin that has come up 'heads' many times in succession is the more likely to come up 'tails') and the hot-hand fallacy (believing that basketball players who have made ten free throws in succession are especially likely to make the next). In the next stage, investors may make decision errors when they buy and sell assets, even when their beliefs are realistic. There are numerous examples, including hyperbolic (or quasi-hyperbolic) discounting, the disposition effect, and the sunk-cost fallacy.

It is often tempting to explain financial market anomalies simply by pointing to one or more of these biases and errors. But such explanations are incomplete and potentially erroneous. One problem is that there are so many documented biases and errors; indeed, a complete list seems not to exist. Given any market anomaly A, a diligent student can always find some decision error or bias B that superficially seems connected, whether or not B really causes A. Even more important, investors' biases and decision errors never translate directly into financial market imperfections. Asset prices are non-trivial functions of investors' buy and sell orders, and they provide information that affects subsequent orders and prices. These later stages of the process depend on the market format, and they can attenuate or amplify investors' biases and errors.

Attenuating Biases and Errors

Three different market forces can greatly attenuate the financial market impact of erratic investors. First, it is a powerful learning experience to lose money in a financial market, or even to see other investors do better when they have no informational advantage. Friedman (1998) and later studies demonstrate that people can overcome even the strongest biases and errors in a suitable learning environment. To the extent that a bias or error leads to clearly inferior performance, an investor will learn to do better over time. Subjects in most laboratory financial markets commit fewer errors and trade more efficiently in later periods than in earlier periods, and subjects with previous experience in a particular laboratory market do better yet.

Second, the market shares of investors with inferior trading strategies tend to shrink over time, reducing their influence on market performance. Blume and Easley (1992) demonstrate theoretically that wealth redistribution eventually eliminates all but the most effective investors. Laboratory studies routinely cancel out this force via stationary repetition, but it can easily be inferred by compounding relative profits across periods.

Third, persistent costly errors and biases create profit opportunities for entrepreneurs whose efforts attenuate (or even eliminate) the market impact. For example, yellow pages and speed dials help us overcome our cognitive limitations in remembering phone numbers. Similarly, mutual funds and a host of investor advisory services allow investors to sidestep their personal biases. Such entrepreneurs can create new problems but, as noted below, those problems also can be studied in the laboratory. Arbitrage is the most direct form of such entrepreneurship. If error-prone investors create an asset price discrepancy, this will attract profit-seeking arbitrageurs whose buy and sell orders tend to make it disappear.

Laboratory studies, including those of Plott and Sunder (1982), confirm the power of arbitrage.

Amplifying Biases and Errors

There are also three strong forces that can amplify the market impact of errant investors. First, raw information is often gathered, analysed and released by individuals who have major personal stakes in the market reaction. Despite oversight by authorities such as the US Securities and Exchange Commission, these individuals may use their discretion to distort the market reaction. Bloomfield and O'Hara (1999) and subsequent laboratory studies confirm the possibility.

Second, professional fund managers typically are compensated (directly or indirectly, via competing job offers) for returns that rank highly relative to their peers. It is difficult to infer from field data whether such incentives have an impact, but inference is straightforward in the laboratory. James and Isaac (2000) find major distortions of laboratory asset prices when traders have rank-based performance incentives, and the distortions disappear in otherwise identical markets when traders are paid only their own realized returns.

Third, and most intriguingly, investors may go astray when they try to glean information from the trades of informed investors. Information mirages (for example, Camerer and Weigelt 1991) can arise as follows. Uninformed trader A observes trader B attempting to buy (due to some slight cognitive bias, say) and mistakenly infers that B has favorable inside information. Then A tries to buy. Now trader C infers that A (or B) is an insider and tries to mimic their trades. Other traders follow, creating a price bubble.

Several research teams (including the author's) have occasionally observed such episodes in the laboratory. They cannot be produced consistently, because incurred losses teach traders to be cautious when they suspect the presence of better-informed traders. The lesson does not necessarily improve market efficiency, since excessive caution impedes information aggregation.

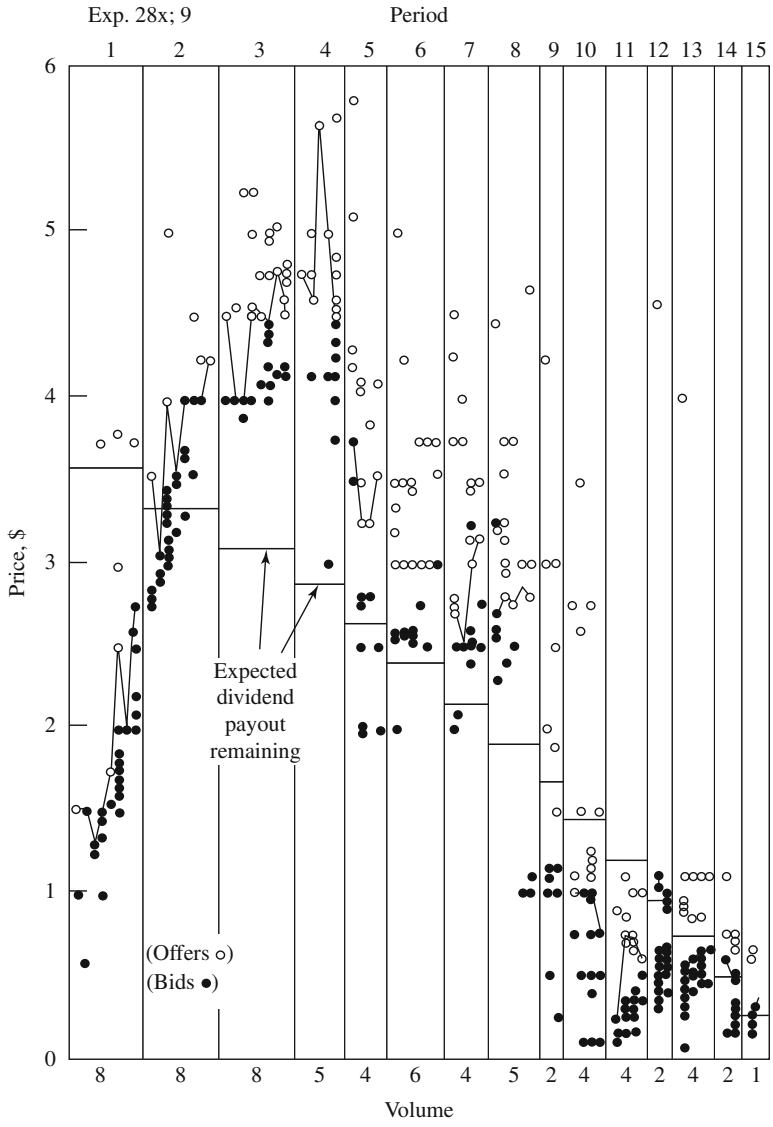
Price bubbles deserve longer discussion, as bubbles have produced important distortions in market prices. Asset prices seemed to disconnect from fundamental value in Japan in the late 1980s, in the dot.com bubble and crash of 1997–2002, and in a number of other episodes since the famous 17th and 18th century events now known as tulipmania and the South Sea bubble. Do such episodes indicate dysfunctional financial markets? Perhaps, but the field data also can be interpreted merely as unusual movements in fundamental value (Garber 1989). By contrast, in the laboratory the experimenter can always observe (or more typically, control) the fundamental value, so bubbles can be detected and measured precisely.

Smith et al. (1988) found large positive bubbles, and subsequent crashes, for long-lived laboratory assets and inexperienced traders. Figure 1 shows a representative example. The expected dividend is constant, so the fundamental value (the sum of expected remaining dividends) declines steadily over the 15 trading periods. Ask ('offer') and bid prices start low, but by the second period the transaction prices (indicated by lines connecting accepted bids and asks) rise above fundamental value. The bubble inflates rapidly until late in period 4. In period 9, prices crash below fundamental value.

Keynes's 'greater fool' theory provides a possible interpretation. Traders who themselves have no cognitive bias might be willing to buy at a price above fundamental value because they expect to sell later at even higher prices to other traders dazzled by rising prices. Subsequent studies confirm that such dazzled traders do exist, and that bubbles are more prevalent when traders are less experienced (individually and as a group), have larger cash endowments, and have less conclusive information.

Current Frontiers: Market Formats, Agents, and Prediction Markets

Which underlying biases and errors are most important? When does attenuation predominate, and when does amplification? Accumulating laboratory evidence inspires new theoretical and empirical field work as well as follow-up laboratory studies.



Laboratory Financial Markets, Fig. 1 A bubble and crash in the laboratory (*Source*: Smith et al. (1988, Figure 9))

It is increasingly clear that answers hinge on the market format or institution – the rules that transform bids and asks into transactions. In particular, the CDA format allows all traders to observe other traders' attempts to buy and sell in real time, and thereby encourages information dissemination. The CDA format attenuates the impact of erratic traders because the closing price is not set by the most biased trader or even by a random trader. The most optimistic traders buy (or already hold) and the most pessimistic traders sell (or never held) the asset, so the closing price reflects the moderate expectations of marginal traders (see section "▶ Smith, Vernon (Born 1927)").

Other traditional formats include the call market (CM), in which bids and asks (or limit orders) are gathered and executed simultaneously at a uniform price, and the posted offer (PO), in which one side (usually sellers) simultaneously announces prices and the other side (buyers) choose transaction quantities at the given prices. Many other formats and hybrids are possible in the Internet age. Which formats are most efficient? Which can attract market share from other formats? Work so far indicates that the CM format does relatively well for thinly

traded assets and the PO format works best when the posting side is more concentrated; but the questions remain far from settled.

Related new work blurs the line between computer simulations and laboratory markets. Computer algorithms for artificial agents, or bots, incorporate specified cognitive limitations, and simulations examine the market level impact (for example, Arthur et al. 1997). Gode and Sunder (1993) showed that simple perishables CDA markets are quite efficient even when populated by zero intelligence (ZI) agents, bots that are constrained not to take losses but are otherwise quite random. Current work puts ZI and more intelligent bots into the same asset markets as human traders, and compares efficiency and the distribution of surplus. Such work should help inform regulators, reformers, and entrepreneurs creating new asset markets. Early published examples of policy-oriented research includes performance assessment of (*a*) trader privileges such as price posting and access to order flow information (for example, Friedman 1993), and (*b*) transaction taxes, price change limits and trading suspensions intended (typically ineffectively) to mitigate price bubbles and panics (for example, Coursey and Dyl 1990).

Prediction markets, which use the information-aggregation property of markets to forecast events such as election outcomes, are gaining increased attention. The Iowa Electronic Market, designed and operated by experimental economists (Berg et al. 2008), offers various assets that pay the holder ten dollars if (and only if) a specified event occurs by a specified date. Participants self-select, are not representative of the general public, and their trades exhibit partisan bias – for example, self-styled Democrats are more likely to buy assets that pay off when the Democratic Party candidates win. Nevertheless, political event asset prices have consistently outperformed opinion polls and all other available predictors. Prediction markets are a growing presence on the Internet, for example tradesports.com, and some corporations such as HP are beginning to rely on them when making business decisions. The line between laboratory and field financial markets is beginning to blur.

See Also

- ▶ Behavioural Finance
- ▶ Smith, Vernon (Born 1927)

Bibliography

Arthur, W.B., J.H. Holland, B. LeBaron, R. Palmer, and P. Taylor. 1997. Asset pricing under endogenous expectations in an artificial stock market. In *The economy as an evolving complex system II*, ed. W.B. Arthur, S.N. Durlauf, and D.A. Lane. Reading: Addison-Wesley.

Berg, J., R. Forsythe, F. Nelson, and T. Rietz. 2008. Results from a dozen years of election futures markets research. In *Handbook of experimental economics results*, ed. C. Plott and V. Smith. Amsterdam: North-Holland (forthcoming).

Bloomfield, R.J., and M. O'Hara. 1999. Market transparency: who wins and who loses? *Review of Financial Studies* 12: 5–35.

Blume, L., and K. Easley. 1992. Evolution and market behavior. *Journal of Economic Theory* 58: 9–40.

Camerer, C., and K. Weigelt. 1991. Information mirages in experimental asset markets. *The Journal of Business* 64: 463–493.

Coursey, D.L., and E.A. Dyl. 1990. Price limits, trading suspension, and the adjustment of prices to new information. *Review of Futures Markets* 9: 343–360.

Friedman, D. 1993. How trading institutions affect financial market performance: some laboratory evidence. *Economic Inquiry* 31: 410–435.

Friedman, D. 1998. Monty Hall's three doors: construction and deconstruction of a choice anomaly. *American Economic Review* 88: 933–946.

Garber, P.M. 1989. Tulipmania. *Journal of Political Economy* 97: 535–560.

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.

Hayek, F.A. 1945. The use of knowledge in society. *American Economic Review* 35: 519–530.

Holt, C.A. 1999. Y2 K bibliography of experimental economics and social science asset market experiments. Online. Available at http://people.virginia.edu/Bcah2k/assety2k.htm. Accessed 19 Feb 2007.

James, D., and R.M. Isaac. 2000. Asset markets: how they are affected by tournament incentives for individuals. *American Economic Review* 90: 995–1004.

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.

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.

Labour Discipline

Peter Hans Matthews

Abstract

Although economists in different fields, or from different schools, use different words to describe the phenomenon, there is widespread agreement that workers can, and sometimes do, 'contest' the sale of their labour power to employers. The question of how employers maintain 'labour discipline' in such an environment has intrigued economists since at least Marx's time.

Keywords

Asymmetric information; Capitalism; Contested transactions; Effort extraction problem; Employment rent; Hobbes, T.; Incentive compatibility; Kalecki, M.; Labour discipline; Labour market contracts; Loyalty; Marx, K.; Supervision; Unemployment

JEL Classifications

J2; J3; J5

Because it is difficult to write and enforce complete contracts in labour markets, transactions are often 'contested' (Bowles and Gintis 1993) and labour discipline must somehow be enforced.

Recent formalizations of the 'effort extraction problem', for example, are premised on the notion that it is difficult for firms to monitor the effort levels of all workers at all times. How much effort workers expend will then depend on, among other things, the cost of job loss. It follows that, as the unemployment rate or, to be more precise, the expected duration of unemployment decreases, the wage at which workers will expend a particular effort level will increase. In many such models, the 'employment rent' consistent with near-full employment is not feasible, and it is equilibrium unemployment that 'solves' the labour discipline problem.

To fix ideas, consider a discrete time variant of the influential Shapiro and Stiglitz (1984) model. There are N identical, infinite-lived and risk-neutral workers, each of whom maximizes the expected value of $\sum_{i=0}^{\infty} \theta^i u(w_i, e_i)$, where:

$$u(w_i, e_i) = \begin{cases} w_i - e_i & \text{if the worker is employed in period } i \\ \overline{w} & \text{if the worker is unemployed in period } i \end{cases}$$

and where wi and ei are the real wage and effort level in period i, θ is the common rate of time preference and is $\overline{w}$ an unemployment benefit, financed, for the sake of convenience, with a lump-sum tax on profits. Workers must choose one of two effort levels, 0 or $\overline{e}$, each period, and there is some likelihood d that a worker who expends no effort in a particular period will be detected and then dismissed. Furthermore, at the end of each period a fraction q of all employed workers enters the jobless pool for other reasons. In a stationary equilibrium, the lifetime utility, V_1, of an employed worker who expends $\overline{e}$ each period will be:

$$V_1 = \frac{w - \overline{e} + q\theta V_3}{1 - \theta(1 - q)}$$

where V_3 is the lifetime utility of a worker who is currently unemployed. (The worker receives $w - \overline{e} + \theta V_3$ and $w - \overline{e} + \theta V_1$ with likelihoods q and $1 - q$, respectively, which implies that $V_1 = q(w - \overline{e} + \theta V_3) + (1 - q)(w - \overline{e} + \theta V_1)$.) In a similar vein, the lifetime utility, V_2, of an employed worker who expends no effort each period will be:

$$V_2 = \frac{w + (d + q(1 - d)\theta V_3)}{1 - \theta(1 - q)(1 - d)}$$

L

Workers will therefore not expend effort $\overline{e}$ unless $V_1 \geq V_2$ or, after substitution and simplification:

$$w \geq \left(\frac{1-\theta(1-q)(1-d)}{\theta(1-q)d}\right)\overline{e} + (1-\theta)V_3 \quad (1)$$

Consistent with intuition, firms will find it more expensive to achieve labour discipline (that is, the incentive-compatible wage will be higher) the costlier effort is to workers, whether this is because the required effort level $\overline{e}$ has increased or the disutility of such effort has. Discipline will also be more expensive when either the likelihood of detection d or the discount rate θ is lower. When, for example, workers care less about the future, the prospect of eventual dismissal will be less salient. An increase in the separation rate q also causes the threshold in (1) to rise: as labour markets become more turbulent, workers have less incentive, *ceteris paribus*, to invest in a particular employment relationship.

To understand the full implications of (1), however, the lifetime utility of unemployed workers must be further decomposed. If a is the fraction of the jobless pool that is (re)hired at the start of each period in equilibrium, the value of V_3 will be:

$$V_3 = \frac{(1-a)\overline{w} + aV_1}{1-\theta(1-a)}$$

when employed workers find it in their interest to expend effort. It is then tedious but not difficult to show that (1) can be written:

$$w \geq \overline{w} + \left(\frac{1-\theta(1-a)(1-q)(1-d)}{\theta(1-a)(1-q)d}\right)\overline{e} \quad (2)$$

In a provocative choice of words, Shapiro and Stiglitz (1984) called this now familiar incentive constraint the 'no shirking condition'. As the likelihood of rehire a tends toward 1, labour discipline becomes impossible to achieve because the incentive-compatible real wage increases without limit. In more intuitive terms, workers are certain to 'contest the exchange' if the expected duration of unemployment, in this case $\frac{1-a}{a}$, and therefore the punishment value of dismissal, are small.

This model and the dozens, perhaps hundreds, of subsequent variations are sometimes viewed as mainstream restatements of the radical position that persistent joblessness is a characteristic feature of capitalism. In Volume I of *Capital*, for example, Karl Marx (1867, p. 701) saw the 'industrial reserve army of the unemployed' as a 'condition of existence of the capitalist mode of production', one which '[held the] pretentions of the active labor army in check' in 'periods of over-production and paroxysm'. Writing almost 80 years later, at the dawn of the Keynesian Revolution, Michal Kalecki (1943, p. 326) would claim that capitalists were 'consistently opposed to creating employment by subsidizing consumption', even if meant a reduction in profits, so that 'discipline in the factories' could be preserved.

The similarities should not be overstated, however. For Bowles (1985), for example, the difference between 'Marxian' and 'neo-Hobbesian' models is the difference between those in which the nature of capitalist production is central and those in which simple 'malfeasance' is the issue. Furthermore, while there is no doubt that Marx believed that the reserve army served to constrain the demands of workers, its existence owes more to the dynamics of accumulation and technological change than to asymmetric information. And, unlike Shapiro and Stiglitz, or for that matter Marx, Kalecki believed the impediments to full employment were largely political, not economic.

The enforcement of labour discipline involves more than reserve armies, however. Levine (1989), for example, extends the Shapiro–Stiglitz model to show that, when firms cannot be sure that low output is the result of low effort, dismissal policies will violate the just-cause principle, and that the (forced) adoption of this principle leads to more efficient outcomes. In other contributions to the literature, enforcement is more subtle. The slope of the representative wage-tenure profile, for example,

which some labour economists believe is too steep to be explained in terms of human capital accumulation alone, could also reflect firms' pursuit of labour discipline: in this case, deferred compensation mimics the properties of a performance bond, and so increases the cost of job loss for recently hired workers.

The substantial variation in the ratio of supervisory to production workers across otherwise similar economies (and over time, for that matter) hints that, in practice, firms can influence the likelihood of detection or, in broader terms, decide how much, and in what form, workers will be monitored. Furthermore, there is reason to believe that, from an efficiency standpoint, firms will spend too much on supervision: if the size of the employment rent were increased at the expense of supervision, the same output could be produced with fewer inputs.

Both the choice of technique and the search for new methods of production influence, and are influenced by, the enforcement of discipline. In some cases, the most salient characteristic of a particular innovation is its effect on effort extraction. As the historian E. P. Thompson (1967) reminds us, for example, the spread of reliable mechanical clocks in production more than two centuries ago represented a watershed in the evolution of enforcement mechanisms, in much the same sense, perhaps, that computerization has, whatever its other effects, forever altered the power to monitor.

Braverman (1974) and others follow this line even further, arguing, in effect, that the widespread adoption of methods of mass production – in particular, the routinization of labour – owed much to how these methods simplified the extraction of effort and reduced replacement costs for dismissed workers. Even if mainstream economists are sceptical, few doubt that the 'rise of the factory' involved 'substantial investment in fixed capital with strict supervision and rigid discipline' (Mokyr 2002, p. 2).

Finally, recent advances in behavioural and experimental economics have revitalized interest in 'bureaucratic control' (Edwards 1977) of the workplace, in which the means to achieve labour discipline are often more subtle. There is considerable experimental evidence, for example, to support the view that workers and firms sometimes exchange 'gifts' of effort and wages, and that this relationship is 'socially embedded' (Gachter and Fehr 2002), one consequence of which is that intrinsic motivation (a sense of loyalty, for example) can also contribute to labour discipline.

See Also

- ▶ Kalecki, Michal (1899–1970)
- ▶ Labour Market Institutions
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Moral Hazard
- ▶ Underemployment Equilibria

Bibliography

Bowles, S. 1985. The production process in a competitive economy: Walrasian, neo-Hobbesian and Marxian models. *American Economic Review* 75: 16–36.

Bowles, S., and H. Gintis. 1993. The revenge of homo economicus: Contested exchange and the revival of political economy. *Journal of Economic Perspectives* 7(1): 83–102.

Braverman, H. 1974. *Labor and monopoly capital*. New York: Monthly Review Press.

Edwards, R. 1977. *Contested terrain*. New York: Basic Books.

Gachter, S., and E. Fehr. 2002. Fairness in labor markets: A survey of experimental results. In *Surveys in experimental economics*, ed. F. Bolle and M. Lehmann-Waffenschmidt. Heidelberg: Physica-Verlag.

Kalecki, M. 1943. The political aspects of full employment. *Political Quarterly* 14: 322–331.

Levine, D. 1989. Just-cause employment policies when unemployment is a worker discipline device. *American Economic Review* 79: 902–905.

Marx, K. 1867. *Capital: A critique of political economy*, 1906. New York: The Modern Library.

Mokyr, J. 2002. The rise and fall of the factory system: Technology, firms and households since the Industrial Revolution. *Carnegie-Rochester Conference Series on Public Policy* 55: 1–55.

Shapiro, C., and J. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74: 433–444.

Thompson, E. 1967. Time, work-discipline and industrial capitalism. *Past and Present* 38: 56–97.

Labour Economics

Richard B. Freeman

Keywords

Affirmative action; Becker, G.; Comparable worth; Decentralized wage-setting; Dual labour markets; Efficient contracts; Elasticities of complementarity; Elasticity of substitution; Forecasting; Harris–Todaro hypothesis; Human capital; Immigration and the city; Internal labour markets; Labour demand; Labour economics; Labour market institutions; Labour supply; Layoffs; Long-term employment; Mincer, J.; Minimum wages; Negative income tax; Period of production; Personnel economics; Profit sharing; Search models of unemployment; Trade unions; Unemployment; Wage differentials; Wage dispersion; Wage rigidity

JEL Classifications

J0

Labour economics studies the demand and supply for the most important factor of production, human beings. Since the days of Marshall and indeed of Smith, if not earlier, economists have recognized that one cannot analyse the market for labour, without taking account of such issues as social relations of production, long-term contractual arrangements, problems of effort and motivation, as well as institutions like unions and internal labour markets, which differentiate the labour market from a bourse. For many years recognition of these factors made labour economics an area in which economic theory was applied sparingly and in which institutional analyses dominated.

This is no longer the case. Sparked in part by theoretical advances and in part by the availability of computerized data-sets with observations on hundreds, (thousands, tens of thousands) of individuals, labour economics underwent a dramatic revolution beginning in the 1960s and accelerating thereafter. As a result modern labour economics diverges notably from its past in two respects: creative use of theory to cast light on the aforementioned aspects of reality and detailed empirical investigations of the behaviour of individuals using advanced econometrics. In addition, in contrast to earlier labour economics, which dealt largely with firms' behaviour from a demand perspective, there has been a pronounced interest in labour supply issues in much of the modern work.

Human Capital

Conceptually the most important development in the rise of modern labour economics has been the '*human capital*' revolution associated with Gary Becker and Jacob Mincer, among others. Human capital analyses concentrate on individual decision-making, particularly with respect to labour supply and related areas of behaviour often associated with sociology rather than economics. Prior to Becker's *Human Capital*, many labour economists tended to regard labour supply decisions as being only loosely based on economic rationality and therefore as a poor subject area for rigorous theory and analysis. By putting decisions regarding education and other forms of improving skills in an investment framework and developing implications for wages, time worked, and diverse other forms of behaviour, the human capital analysis fundamentally changed the way in which economists see labour supply. The simple investment concept – that individuals, like enterprises, 'invest' early in life (through schooling, and on-the-job-training) and reap rewards later, thereby producing an upward tilt to the age-earnings profile – has proved valuable in interpreting wages, and in directing attention to lifetime considerations in labour supply (for example, use of deferred compensation to motivate workers). Equally important, the view that diverse forms of decision-making can be fruitfully analysed by economic models of rational behaviour has illuminated not only traditional areas of labour supply behaviour such as labour participation, hours worked, job search, career choice, and

the like, but has also extended the boundary of analysis to issues ranging from crime to marriage, fertility, and health.

At roughly the same time that human capital theory directed attention at individual behaviour, *computerized data-sets* providing information on the economic and demographic characteristics of individuals became available to analysts. The conjunction of theory and data produced a massive outpouring of studies on the effect of individual as opposed to market or employer factors on wages, and on the supply decisions of individuals. As a result of these factors the labour economist of the 1980s differed substantively in his or her orientation and analytic approach from the labour economist of earlier decades. Whereas in the 1950s labour economists generally studied wages and mobility at the level of industry, area, or in some cases establishments, in the 1970s and 1980s they tended to focus on individuals, first with cross-sectional data comparing different people, then with longitudinal (or panel) data that follow the same person over time. Whereas in the 1950s labour economics was heavily concerned with case studies, in the 1970s and 1980s labour economics had become pre-eminently the field of applied econometrics and statistical analyses of large data types.

In addition to use of modern theoretical and econometric tools, labour economics had been intimately involved in development and analysis of 'controlled experiments' to explore labour supply responses to alternative tax or welfare systems. The most famous of these experiments, the New Jersey and Seattle–Denver experiments, used a control methodology to explore the potential effects of a negative income tax, finding labour supply elasticities that ranged from modest (men) to significant (women) and also uncovering some forms of behaviour relatively hard to explain by standard economics theories (notably in family behaviour). Despite problems with the experimental approach, it marks a striking advance in the set of tools which are employed to explore supply issues.

While there will be some disagreement among economists about the contribution of the human capital and human capital-inspired analysis to explanation of social phenomena, a reasonable assessment is that the analysis has done a good job in illuminating a broad area of social behaviour but at the same time has not explained most of what goes on in the labour market. Changes in behaviour and in structural relations for reasons of tastes, technology, or whatever, create variation at a point in time and changes over time that are not readily explicable by standard models. For example, in the area of female labour participation, studies find that income effects (reflected in husband's income) and substitution effects (reflected in the wages of the woman) and various indicators of the shadow price of time, such as number of young children, have the sorts of impacts on participation one would expect, but that these factors cannot readily account for the magnitude of upward trends in participation or for cross-country differences in trends or levels. Similarly, while the magnitude and probability of punishment and rates of unemployment and related labour market factors affect crime, they do not account for the high rates of crime in the US relative to other countries not for the time series pattern of change in crime in the US.

Even in terms of wage determination, while the variables associated with human capital enter equations with high significance, they are not the dominant factor in variations in wages among individuals: in a typical log-earnings equation, education may explain five per cent of the variation and education and years of experience may explain 15 per cent in total, with job tenure (whose effect is partly the outcome of on-the-job training and partly the result of institutional seniority rules) dominating the experience component; additional important contributors to wage variation include such factors as industry and firm (or establishment) of work that cannot be readily interpreted solely by supply-side factors.

Labour Demand

The theoretical and empirical thrust of modern labour economics has had less impact on analyses and understanding of *demand for labour and firms' behaviour* than it has had on the supply of

labour. One reason is that previous generations of scholars had devoted considerable effort to analysing the demand side, dealing with such issues as internal labour markets, hiring, promotion, and wage policies, and the structure of wages in various markets, yielding a body evidence on behaviour which has stood up to further analysis. Another reason is that cross-section and longitudinal data on firms and establishments comparable to that on individuals have not been readily available. The computerization of personnel records of firms provides the best potential for major empirical advances in analysis of their labour demand and personnel policy, but as yet work on these records has been rather sparse.

The modern analysis of labour demand has taken the key facts established by the previous generation – that labour markets are far from 'spot markets' – and sought to develop a consistent theory of economic behaviour, in which the firm is viewed as choosing a particular wage and personnel policy to optimize its profits, given the likely response of workers to the policy. Since firms will do best if they offer a labour compensation package that workers desire (at a given cost), some analysts look upon the firm as implicitly maximizing the utility of workers. Others pay greater attention to areas of conflict between the two sides, dealing with issues of shirking, (which makes deferred compensation especially valuable) and effort.

Thus far, the success of this approach has been more on the theoretical than empirical front. Analysts have developed models for such phenomena as deferred compensation, piece rates and related 'prize' systems for rewarding workers, and for such policies as mandatory retirement, but the ability of these 'stories' to account for the bulk of observed variation has not been demonstrated. To take one example, these are unquestionable differences in pay among firms to local labour markets: some firms pay what appear to be 'above-market' rates, while others pay less than the going rate. One can tell efficiency wage stories (firms pay high wages to reduce turnover and shirking); rent-sharing stories (firms share their economic rents with workers); or union-threat stories (firms pay to keep unions out) about such policies; but labour economics has yet to determine the relative empirical relevance of these stories. In that sense, progress beyond the work of the generation of the 1940s and 1950s that stressed the firms' wage policies has been limited.

Another area of work on demand, more grounded in the neoclassical model of the firm, has examined the magnitude of elasticities and cross-elasticities of labour demand for workers of different skills and the effect of administered wages (minimum wages) on employment. Since the basic parameters in labour demand analysis are elasticities of demand one would hope that empirical work would pin down their magnitude with some certainty. Such has not always been the case. In the US most studies, including those focused on the minimum wage, yield relatively modest elasticities for low-wage workers and manufacturing labour, usually considerably below unity. Analysis of demand for women workers in Australia, exploiting an exogeneous change in female wages due to comparable-worth-type rulings, has also found relatively moderate demand responses. Work on the UK and some European countries, by contrast, has yielded larger estimates of elasticities, which is puzzling given the widespread belief that the United States has a more flexible labour market with employers able to adjust employment more freely than in Europe.

Analyses of elasticities of substitution (which measure the effect of changes in relative wages on changes in relative employment) and of elasticities of complementarity (which measure the effect of changes in relative employment on relative wages) for narrowly defined skill, age, or education groups tend to find higher elasticities, implying that a large exogenous increase in the relative number of persons in a group can significantly affect the relative wages. Two cases in point are the 1970s increased number of young workers ('baby boomers') and of young college graduates in the United States, which greatly reduced the earnings of those groups relative to older and less educated groups.

As a general rule, shifts in demand schedules tend to account for more observed changes in employment than do movements along demand

schedules. Work on factors shifting demand for labour (technology, changes in consumer tastes, income elasticities for the goods produced by particular groups of workers) has, however, been rather limited. One body of work has focused on relative demand for minorities, where the development of specific programmes to raise demand provides the same sort of exogeneous shift in the curve as minimum wages provide movements along the curves. The available evidence here suggests that affirmative action and similar programmes have played a role in raising demand for minority labour in the United States, though here as elsewhere changes in the market cannot be solely attributed to one demand-shift factor. Another body of work, associated more with governmental agencies than with academic economists, has projected future labour 'requirements' in an input–output framework.

Comparing the theoretical and empirical work on demand, one is struck by the failure of the empirical analysis to take appropriate account of the potential importance of the long-term employment arrangements and internal labour markets stressed in the theory. A major cause of the difficulty is a data problem: until analysts of labour demand have available detailed longitudinal data on employment by establishment or firm, and on firms' personnel and wage policies, it is exceedingly difficult to marry the advances in theory to the data.

The contrast with the supply side, where theory and data came together, highlights the complementarity of the two 'blades' of the research scissors for a field to develop rapidly.

Institutions

In the area of institutions labour economics has tended to focus on unions as the major worker institution in modern capitalist economies. A massive body of work has examined the effects of unions on wages, beginning first with comparisons of wages in union and non-union sectors of the economy (industries, occupations across cities, and so forth), and then moving on to analysis of the computerized data-sets with information on individuals, classified by union status. While the question that motivates this work is 'what do unions do to the economy?' the empirical analyses have, of necessity, been devoted to measuring differences between union and nonunion workers (firms).

Following a massive outburst of work on union–non-union wage differentials, labour economists turned to a wide variety of behaviour by individuals and firms likely to be affected by unions. Analysts found quits to be lower and job tenure higher under unionism; temporary layoffs (which occur when workers are laid off for short periods of time, then recalled) to be largely a union sector phenomenon: and the dispersion of wages to be lower in union plants, as well as finding effects of unions on profitability and productivity. This work has paralleled the human capital analysis by continually expanding the set of outcome variables under study and the labour demand analysis by focusing on issues dealt with by the earlier generation of labour economists.

On the theoretical front the thrust of modern work on unions has explored the idea of 'efficient contracts' in which unions and management eliminate potential inefficiencies due to monopoly through joint wage and employment determination. Efforts have also been made to develop models of unions as maximizing institutions, following the path laid out by Dunlop in the 1940s, in which unions are concerned with both wages and membership or job security.

Markets

Demand, supply and institutions interact in market settings, and labour economics contains numerous studies of the operation of labour markets for various types of labour. Attention has shifted from markets for blue-collar labour to markets for white-collar labour, and from case studied to more econometric investigations of wage, employment, and unemployment.

One strand of work, closely related to human capital analysis, has been to investigate markets for highly educated workers, where the time period of 'production' (college takes four years)

allows one to differentiate supply and demand forces in the market. The first generation of such models used relatively simple cobweb structures; a later generation examined more complex rational expectations market-clearing models. The general tone of the results has been sufficiently successful to change the issue from whether markets follow readily understandable economic principles to which type of model best explains patterns of change. Even so, here as elsewhere in economics, the models have not done an especially good job in forecasting, in large because of our inability to project *shifts* in demand schedules, noted earlier.

Another stream of market analysis had dealt with such topics as geographic and industrial mobility, and unemployment and related wage patterns. Observed patterns of wages and mobility make it clear that in the United States decentralized wage setting across a huge geographic area produces separate local labour markets which experience different patterns of change, with costs of mobility sufficiently large as to produce significant 'losses' to some (particularly older) displaced workers. An important empirical finding has been that high-wage cities tend to have high unemployment, providing some support for 'job search' as a factor in unemployment. Across industries, the United States evidence shows falling dispersion in wages in periods of economic boom (as low-wage employers raise pay while high-wage employers do not) and also an upward trend in dispersion of wages among industries. Other countries do not appear to have experienced such a trend over time, possibly because of centralized wage setting.

The question of whether unemployment is a long-term or transitory phenomenon has been analysed in the context of models which differentiate between completed and uncompleted spells and between the duration and incidence of unemployment. Perhaps the most important finding, which appears to hold for a large number of countries, is that the bulk of unemployment at any one time is due to a small number of people who are unemployed for long periods, rather than to short-term unemployed people.

Finally, an important area of labour research which diverges substantively from the micro-orientation of much of modern labour economics has involved analysis of macro-change in wages, employment and unemployment over time within a country and across countries. To some extent, labour economics has played a 'devil's advocate' role with respect to proposed macro-explanations of problems like unemployment and wage inflation. Macroeconomists have suggested that unemployment is due to such factors as rigid wages associated with three-year contract cycles, intertemporal substitution of time, shocks that require mobility across sectors; labour economists have tested and, in general, rejected these models in a macro-context.

In addition, however, studies suggest that different labour market institutions in different countries may affect macro-outcomes as well. An important hypothesis has been that 'corporatist' or centralized free/market economies have an advantage in adjusting to stagflation because all workers can jointly agree to lower rates of increases in wages, avoiding Prisoner's Dilemma problems. Another hypothesis has been that 'flexibility' in labour markets is the key to the differential performances of the European and the American economy in employment generation in the 1970s and through the mid-1980s. In the area of theory the notion that 'a share economy' (where workers are paid in part via profit or revenue sharing) may produce less unemployment than a 'wage economy' has directed attention at alternative modes of paying workers, particularly over the business cycle. Whether comparative analysis focusing on different wage-setting mechanisms across countries becomes a major part of the field, however, remains to be seen.

Another area of comparative labour market studies that proliferated in the 1960s and 1970s focused on labour markets in developing countries. The Harris–Todaro model, which interpreted urban unemployment in terms of migration to cities and queuing for high-wage jobs, directed attention at mobility issues and institutional forces causing 'dual labour markets'. A variety of studies dealing with the effect of education and human capital on earnings and behaviour revealed

patterns similar to those in developed lands, suggesting that some aspects of markets function similarly across levels of development.

Conclusion

In the span of two decades labour economics has moved from a largely institutional field into the mainstream of economics, while maintaining its empirical bent. It has widened the subject of discourse, particularly on the supply side, and struggled to synthesize the 'facts' of the labour market with economic principles. It is the interplay of detailed micro data and economic analysis which currently is the hallmark of the field, differentiating it from more abstract theoretical and less factually based parts of the discipline.

See Also

- ▶ Human Capital
- ▶ Industrial Relations
- ▶ Labour Economics (New Perspectives)
- ▶ Strikes
- ▶ Women's Work and Wages

Bibliography

Abraham, K., and J. Medoff. 1980. Experience, performance, and earnings. *Quarterly Journal of Economics* 95: 703–736.

Ashenfelter, O. 1984. *Macroeconomic analyses and microeconomic analyses of labour supply*, Working paper no. 1500. Cambridge, MA: NBER.

Ashenfelter, O., and J. Heckman. 1974. The estimation of income and substitution effects in a model of family labour supply. *Econometrica* 42: 73–85.

Ashenfelter, O., and R. Layard, eds. 1984. *Handbook of labour economics*. Amsterdam: North-Holland.

Becker, G. 1964. *Human capital*. New York: Columbia University Press for the NBER.

Becker, G. 1976. *The economic approach to human behavior*. Chicago: University of Chicago Press.

Becker, G. 1981. *A treatise on the family*. Cambridge, MA: Harvard University Press.

Brown, C., C. Gilroy, and A. Cohen. 1982. The effect of the minimum wage on employment and unemployment: A survey. *Journal of Economic Literature* 20: 487–528.

Bruno, M., and J. Sachs. 1985. *Economics of world wide stagflation*. Cambridge, MA: Harvard University Press.

Clark, K., and L. Summers. 1979. Labor market dynamics and unemployment: A reconsideration. *Brookings Papers on Economic Activity* 1979 (1): 13–72.

Doeringer, P., and M. Piore. 1971. *Internal labor markets and manpower analysis*. Lexington: Heath.

Ellwood, D. 1986. The spatial mismatch hypothesis are there teenage jobs missing in the ghetto? In *The black youth job crisis*, ed. R. Freeman and H. Holzer. Chicago: Chicago University Press.

Farber, H. 1984. The analysis of union behavior. In *Handbook of labor economics.* Amsterdam: North-Holland.

Freeman, R. 1971. *The market for college-trained manpower*. Cambridge, MA: Harvard University Press.

Freeman, R. 1983. Crime and unemployment. In *Crime and public policy*, ed. J. Wilson. San Francisco: Institute for Contemporary Studies.

Freeman, R., and J. Medoff. 1984. *What do unions do?* New York: Basic Books.

Gregory, R.G., and R.C. Duncan. 1981. Segmented labor market theories and the Australian experience of equal pay for women. *Journal of Post Keynesian Economics* 3: 403–428.

Hall, R. 1975. The rigidity of wages and the persistence of unemployment. *Brookings Papers on Economic Activity* 1975 (2): 301–349.

Hamermesh, D., and J. Grant. 1979. Econometric studies of labour–labour substitution and their implications for policy. *Journal of Human Resources* 14: 518–542.

Harris, J.R., and M.P. Todaro. 1970. Migration, unemployment and development: A two sector analysis. *American Economic Review* 60: 126–142.

Hausman, J., and D. Wise. 1985. *Social experimentation*. Chicago: University of Chicago.

Heckman, J. 1974. Life cycle consumption and labor supply. *American Economic Review* 64: 188–194.

Killingsworth, M. 1983. *Labour supply*. Cambridge: Cambridge University Press.

Lazear, E. 1979. Why is there mandatory retirement? *Journal of Political Economy* 87: 1261–1284.

Leonard, J. 1985. The effectiveness of equal employment law and affirmative action regulation, Working paper no. 1745. Cambridge, MA: NBER.

Lewis, H.G. 1963. *Unionism and relative wages in the United States*. Chicago: University of Chicago Press.

Lewis, H.G. 1986. *Union relative wage effects: A survey.* Chicago: University of Chicago Press.

Mincer, J. 1962. Labor force participation of married women. In *Aspects of labor economics*, ed. J. Mincer. Princeton: Princeton University Press.

Mincer, J. 1968. Labor force participation. In *International encyclopedia of the social sciences*, vol. 8. New York: Macmillan.

Rees, A. 1962. *The economics of trade unions*. Chicago: University of Chicago Press.

Rosen, S. 1984. Distribution of prizes in a match-play tournament with single eliminations, Working paper no. 1516. Cambridge, MA: NBER.

Rosen, S. 1985. Implicit contracts: A survey. *Journal of Economic Literature* 23: 1144–1175.

Segal, M. 1986. Post-institutionalism in labor economics: The forties and fifties revisited. *Industrial Labor Relations Review* 39: 388–403.
US Department of Labor. 1985. Projections of the economy, labor force, industrial and occupational change to 1995. *Monthly labor review*, November.
Watts, H., and A. Rees, eds. 1978. *The New Jersey income maintenance experiment*. New York: Academic Press.
Weitzman, M. 1985. *The share economy*. Cambridge, MA: Harvard University Press.

Labour Economics (New Perspectives)

Christopher Taber and Bruce A. Weinberg

Abstract

Since Richard Freeman wrote labour economics for the first (1987) edition of *The New Palgrave: A Dictionary of Economics,* labour economics has become increasingly empirical, with less emphasis on theory. The most noticeable change in empirical work is an increased emphasis on the plausibility of identification assumptions such as the validity of instrumental variables. Among the areas growing or receiving the greatest attention are changes in the wage structure, the economics of education, social interactions and personnel economics. The range of topics studied by labour economists today has broadened far beyond those of traditional labour economics.

Keywords

Education production functions; Fixed effects; Group selection; Human capital; Identification; Instrumental variables; Labour economics; Labour market search; Matching; Natural experiments; Personnel economics; Returns to schooling; Roy model; Sample selection problem; Skill-biased technical change; Wage differentials; Wage inequality, changes in

JEL Classification

J2

When Richard Freeman wrote his excellent article labour economics for the first (1987) edition of *The New Palgrave: A Dictionary of Economics,* which is reproduced in the present edition, labour economics had changed dramatically with the development of the human capital paradigm and the use of large-scale data-sets. In many ways labour economics has continued along the trends Freeman discussed, but in other important ways its focus has shifted, in terms of both topics and interests. The goal of this article is to describe major trends in this dynamic field of applied microeconomics since the 1980s. We begin with an overview of methodological trends that are common to much of the field and then talk about specific research questions within labour economics. We will direct readers to the appropriate *New Palgrave* articles for a more complete discussion of those topics.

One important way in which labour economics has changed since Freeman's article is that it has become increasingly empirical. Presumably, this trend is due at least in part to improvements in large-scale computing and ease of access to data sources. Along with this trend has come a decreased emphasis on theory in all but a few areas. The decreased emphasis on institutional factors, discussed in labour economics, has certainly continued (even the study of labour unions has declined substantially). Along with the trend towards increased empirical work has come a much stronger emphasis on the plausibility of identification assumptions. In many labour contexts, there are substantial unexplained variation in the dependent variables being studied, leading to interest in strategies for dealing with sample selection and endogeneity. In the case of earnings regressions, for instance, the vast majority of the variation in earnings cannot be explained by observable worker characteristics. While the presence of important unmeasured factors does not invalidate a regression model, it raises a concern that the coefficients on the variables of interest may be biased if the substantial unexplained component is correlated with the variables of interest. In the earnings regression case, one worries that workers who are more able or motivated (in ways that are unmeasured by the analyst) may obtain

more school, biasing estimates of the return to school upwards. Labour economists have increasingly focused on these selection or endogeneity issues and this emphasis has spilled over from labour economics into other fields in economics.

The most noticeable change in approach is much greater emphasis on the plausibility of identification strategies. The two most noticeable examples in this vein are an increased use of fixed effect approaches (including 'difference in differences') and much more attention being paid to the validity of instrumental variables. A classic example is Angrist's (1990) study of Vietnam veterans. Estimating the effect of veteran status on earnings is plagued by the classic sample selection problem. Angrist solves this problem by using the Vietnam draft lottery number as an instrument for veteran status. This number is mechanically related to veteran status, but since it is random it will be unrelated to earnings again by construction. Studies along these lines are typically referred to as experimental or natural-experiment studies (depending on whether the variation arises from an explicit randomized experiment or policy or institutional factors that are plausibly, but not explicitly, random).

Structural estimation has also received substantial attention since the 1980s, although in relative terms, substantially less than during the previous 20 years. Different people may define structural in different ways. For example, simple linear models estimated by ordinary least squares or two-stage least squares can be considered structural if the researcher is explicit about the interpretation of the parameters. We have witnessed a large increase in popularity of a more ambitious approach in which a researcher formally models an individual's decision process and estimates the underlying parameters of say the utility function or production process by choosing the parameters that minimize the difference between observed outcomes and those implied by the model. For instance, a young individual has the option to attend school, or work in a variety of jobs, or remain in the household sector in each year of his or her life. One structural approach would be to estimate an individual's value function from the terminal period backwards at each node on the decision tree by matching observed behaviours to those implied by utility maximization. Unobserved individual factors can be addressed by including them in the value functions and integrating them out when trying to match the data. This approach is computationally demanding. Substantial advances in computational methods for these models and improvements in computer technology have allowed researchers to estimate considerably richer models. This approach has benefited from the year-by-year extension of longitudinal (panel) data sets.

The literature on returns to schooling provides a nice example of the evolution of empirical approaches. The goal of this literature is to estimate the causal effect of schooling. Willis and Rosen (1979) is a classic paper in this literature and an excellent illustration of empirical approaches prior to 1987. These authors consider a model with two schooling choices, high school and college, in which students make decisions to maximize the present value of earnings. They allow for individual heterogeneity in college and high-school earnings, college and high-school earnings growth, and interest rates. Their empirical approach consists of a three-stage method in which the first stage is a reduced form probit for college attendance. The second stage is a series of wage regressions including inverse Mills ratios. The third is a 'structural probit' that allows one to estimate the effects of earnings on schooling choices. The key for semiparametric identification in models like this is a variable that affects schooling choices but does not affect earnings directly (see Roy model for discussion of semiparametric identification in this type of model). Willis and Rosen use family background as their exclusion restriction. Family background is relatively strongly related to schooling and might not directly affect earnings. However, subsequent researchers have been sceptical about this exclusion restriction. The biggest concern in using regression analysis is that schooling is probably related to unobservable ability, but for similar reasons one may expect family background to be related to unobserved ability. Either through genetics, parenting skills, or simply resources one might worry that children from privileged

backgrounds have more unobserved ability than their less fortunate peers.

Since 1990 or so many papers have tried to develop more credible exclusion restrictions to estimate the return to schooling. One of the most well known is Angrist and Krueger (1991), who use quarter of birth as an instrument. They argue that a combination of truancy laws and school starting ages will lead students born late in a calendar year to obtain more education than a student born early in a year. To see why, suppose that the cut-off date for starting school is 1 January. As a result, an individual born on 31 December 1962 will begin school a year earlier than a student born a day later, on 1 January 1963. However, if both of these students drop out of high school as soon as the truancy law says that they can, say on their 16th birthday, then the student born in December will have attained an extra year of schooling. Unfortunately, data-sets are not sufficiently large to focus only on these 2 days, so Angrist and Krueger (1991) use quarter of birth instead. Furthermore, there is a fair amount of slippage in that neither truancy laws nor age cut-off dates are strictly adhered to. Note that this last feature does not invalidate the instrument but reduces its power. As an example of a fixed effect approach, Ashenfelter and Krueger (1994) collect data on both earnings and educational attainment of twins. By using family fixed effects they can obtain an estimate of the returns to schooling, differencing out genetic ability.

At the same time we have seen a large structural literature emerge that has generalized Willis and Rosen (1979) by allowing for more complex educational choices and selection. A classic example of this approach is Keane and Wolpin (1994), who estimate a dynamic model of labour-market decisions. They generalize Willis and Rosen (1979) by allowing for many more than two schooling choices (high school versus college), by allowing students to go back and forth from the labour market and school, and by allowing the payoff to schooling to be sector specific. Another example is Heckman et al. (1998a), which estimates a general equilibrium version of the Willis and Rosen (1979) model that estimates not just the pricing equation for schooling but the determinants of both the supply and demand for college. This additional structure included in these papers allows one to simulate substantially more complicated policy experiments than one can perform with the Willis and Rosen (1979) framework.

There are substantial disagreements over the relative merits of different empirical approaches, and a full discussion goes well beyond the scope of this article. With that in mind, some of the instrumental variable and difference in differences approaches have the benefit of placing identification and the source of identification at the forefront of the analysis. For example, the source of identification in the Angrist and Krueger (1991) case is transparent. Estimation of intricate structural models typically requires substantial assumptions, whose validity is frequently unclear, but because the underlying parameters of the problem (preferences, the technology and so on) can be estimated, it is frequently possible to evaluate a wide range of policies that are not represented in the data. For example, identification in Keane and Wolpin (1994) is much less transparent. While it may appear that the reduced-form, natural experiment approach requires fewer or weaker assumptions, work of this type usually implicitly makes a number of important assumptions, particularly if one wants to apply these results to some other context. For example, Heckman et al. (1998b) demonstrate that one can severely underestimate policy effects if one ignores general equilibrium (GE) effects in their model. When researchers ignore GE effects in drawing policy predictions from their work, they implicitly assume that the demand for educated workers is perfectly elastic. The work of Heckman et al. (1998b) suggests that this is a very strong assumption. It seems likely that a large variety of approaches will continue to be used and that results that are robust across a wide range of approaches will be most convincing.

As indicated, labour economics has become increasingly empirical as emphasis on identification has increased. Personnel economics the study of incentives within firms, is a notable exception. This literature is discussed in more detail in

personnel economics. Another exception is work on search and matching, which spans labour economics and macroeconomics and which tends to be more theoretical. Because it requires explicit statements of the decision problem, structural work tends to be more theoretical than reduced form work, although deriving explicit theoretical results is rarely the focus of such studies.

Another notable recent development in labour economics is that the scope of problems that labour economists address has broadened considerably since the 1980s. However one wants to define 'labour economics' – as the study of the determinants of individual earnings, the demand and supply for labour, and the functioning of labour markets or as whatever labour economists do – what is noteworthy is that much of the work being done by labour economists falls well outside a traditional definition of the field. Similarly, other fields have increasingly drawn on ideas developed in labour economics, and the lines between labour economics and closely related fields, including development, urban, and public economics, are blurring. To some extent this reflects the general applicability of traditional labour theory (for example, human capital, which has played an important role in growth economics) and to some extent it reflects the widespread applicability of the econometric techniques developed by labour economists.

One of the most influential areas in labour economics since the 1970s has been the changes in the wage structure (see wage inequality, changes in). Much of this work focuses on the increase in inequality and the increase in the returns to education in the United States. This literature has emphasized demand-side factors, and skill-biased technological change in particular, as the primary explanation for the recent trends. Most recent work on the demand-side of labour economics has been in this area. Whether it is a result of this literature, or coincidental, the whole field has shifted towards trying to understand wage differentials and human capital accumulation.

Another increasingly active area in labour economics is the economics of education, which perhaps can be considered its own field rather than a subfield of labour. The increased interest in education probably has arisen both from the literature on the changing wage structure, which emphasizes human capital, and from the increasing attention to education in the policy world. Within labour economics, understanding the economic value of education is one of the most studied empirical questions (see returns to schooling for a summary of this literature). Economists have also moved from wanting a general understanding of the effects of education on wages to a more specific understanding of what aspects of school are most important in forming human capital. Specifically, researchers have tried to uncover these factors in the 'education production function' literature discussed in education production functions.

We have also seen increased research on private schools and school choice (see school choice and competition for a discussion of this literature). Another branch of this literature (which is really much more of a subfield of public economics rather than labour economics) studies the complicated system under which schools are financed and how changes in these schemes influence students (see educational finance for a description of this literature).

Empirically, race and gender are also important determinants of wages. There is a long literature in labour economics on the economics of discrimination, which tries to understand why these differences arise. The two most studied effects have been the male/female and black/white gap in the United States. While the raw log wage differentials are of similar magnitude (approximately 20%), the effects are very different from each other. As more controls are included in the analysis the black–white gap declines substantially; see black–white labour market inequality in the United States. This has led researchers to focus on pre-market forces as the primary cause. By contrast, men and women look much more similar when they enter the labour market. Thus the difference seems to be related to post-market entry factors. women's work and wages discusses this literature.

The traditional field of labour supply has probably received less attention since the last Palgrave

than in the decades preceding it. Much of the work on this subject has focused on the lower end of the earnings distribution. Perhaps most importantly, a large literature has arisen that attempts to measure the effects of transfer programmes on labour supply of low-income individuals, especially on single mothers. Another important policy area has focused on understanding the effects of minimum wages on employment. Related to the literature on the changing wage structure, there has also been a substantial literature studying labour-force participation among low-skilled workers who are likely to be close to the margin to work and whose wages have fallen considerably (in the case of the United States). While it is well known that labour-force participation among women has increased substantially, participation among men has declined (see for example, Juhn et al. 2002). This literature is discussed more thoroughly in labour supply.

Drawing on research in sociology, labour economists have also become increasingly interested in how people are affected by the groups (for example, schools or neighbourhoods) to which they belong (see social interactions (empirics) and social interactions (theory)). Such studies span a number of the topics already discussed – how students' educational outcomes (and other behaviours such as substance use) depend on those of their peers; or how labour market activity (for example, employment or welfare participation) depends on that of neighbours. Naive estimates indicate that people's behaviours and outcomes are highly correlated with those of their groups, but researchers have been concerned that the groups that people choose (or are 'forced' into) are similar to themselves. A substantial literature has developed using quasi-experiments and explicit experiments to estimate the effect of groups on the people who are in them controlling for the selection processes into groups. Estimates that control for the selection processes are considerably lower than those that do not.

While individual characteristics are very important for wage determination, characteristics of the firm may matter as well. There have been an increasing number of data-sets that allow researchers panels on both firms and workers. These types of data-set allow one to use procedures such as estimating both firm and worker fixed effects (see for example, Abowd et al. 1999). These papers show that firm effects are an important component of wages. The most obvious explanation for this type of result is that there is some type of friction in the labour market. Perhaps as a result there has been an increased interest in labour market friction and its importance in explaining inequality: see labour market search for a discussion of this search literature and matching for a discussion of the matching literature.

There has also been increased attention on the economics of the household, which lies at the intersection of labour economics and other fields such as demography. This work includes studies of bargaining between members of the household on intra-household resource allocations and the effect of household behaviours on children's human capital. Related to these topics are marriage and fertility decisions and household labour supply.

In recent years, the theoretical concepts and empirical methods of labour economics have proven useful across a wide range of topics. Consequently, labour economics has influenced work in a wide variety of other areas and the topics studied by labour economists have expanded considerably.

See Also

- ▶ Labour Economics
- ▶ Personnel Economics
- ▶ Returns to Schooling
- ▶ Roy Model
- ▶ Wage Inequality, Changes in

Bibliography

Abowd, J., H.F. Kramarz, and D. Margolis. 1999. High wage workers and high wage firms. *Econometrica* 67: 251–334.

Angrist, J.D. 1990. Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review* 80: 313–335.

Angrist, J., and A. Krueger. 1991. Does compulsory schooling attendance affect schooling and earnings? *Quarterly Journal of Economics* 106: 979–1014.

Ashenfelter, O., and A. Krueger. 1994. Estimates of the economic return to schooling from a new sample twins. *American Economic Review* 84: 1157–1173.
Heckman, J., L. Lochner, and C. Taber. 1998a. Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics* 1: 1–58.
Heckman, J., L. Lochner, and C. Taber. 1998b. General equilibrium treatment effects. *American Economic Review* 88: 381–386.
Juhn, C., K. Murphy, and R. Topel. 2002. Current unemployment historically contemplated. *Brookings Papers on Economic Activity* 2002(1): 79–136.
Keane, M.P., and K.I. Wolpin. 1994. The career decisions of young men. *Journal of Political Economy* 105: 473–522.
Willis, R.J., and S. Rosen. 1979. Education and self selection. *Journal of Political Economy* 87: S7–S36.

Labour Exchange

S. Olivier, J. Bonar and J. D Rogers

This term is sometimes used loosely as the equivalent of labour registry. Accurately and historically it applies to a class of institutions which found much theoretical favour amongst early cooperators and the associates of Robert Owen's propaganda. Numerous labour exchanges, marts, and banks flourished in England in 1832, 1833, and 1834 for the direct exchange of the products of labour according to the amount of labour expended in making them, without the intervention of money or the expenses of the ordinary machinery of distribution. Their fundamental principle was the doctrine that labour is the source of all wealth, and labour-cost the true measure of value: the operation of this principle was considered to be interfered with distorted by the intervention of money, a monopolized and limited commodity, as a medium and essential of exchange. The exchanges met a popular requirement, and, had the constant efforts of the more clear headed among their directors been successful in maintaining a strict commercial system of valuation, might have been long-lived. But the labour-value theory, and the conventional rating of all labour at sixpence an hour for purposes of valuation in exchange, or for labour notes, defeated these efforts. The valuation of the price of materials was also a constant difficulty. Sharp tradesmen took labour-notes in their shops, and picked out the goods in the exchanges that were saleable at a profit on their 'labour value'. This process accelerated the accumulation of stocks so that no one cared to take at the price of sixpence per hour for the time of their makers: the 'labour-note' became depreciated *pari passu* with this depreciation of the security on which it rested; its depreciation enabled traders who took it to skim the deposits still closer, until the goods in stock, and the labour-note, had fallen to a commercial value below that which the workman of average skill could earn in the ordinary labour market in the time represented by their price, and the exchanges one by one collapsed, after furnishing a very interesting illustration to the history of theories of value.

The history of this movement may be best brought under four heads: I. The Proposal. II. The Scheme. III. Labour Exchange Notes. IV. The Principles on which Labour Exchanges were based.

The Proposal

In 1820 Robert Owen wrote that there were three stages in the history of exchange: (1) barter, which admitted 'the only equitable principle of exchange', which was to exchange 'the only equitable principle of exchange', which was to exchange 'the supposed value of labour in one article against the amount of labour contained in any other article' ('Report to county of Lanark', *Autobiography*, ii. 278). As wealth increased, barter became impossible, and (2) artificial exchange, or exchange through some medium with a value of its own, introduced the commercial stage, which forgot 'the natural standard of labour'. But the increase of wealth was superseding the use of the gold and silver standard, and had partly done so during the suspension of cash payments between 1797 and 1819 (ibid., p. 266).

(3) The third stage began when exchange would be 'equitable' as in the first stage, and by means of a medium, as in the second stage. The new medium, in order to reflect without deflecting the 'natural standard of value', should not possess a value of its own. What was it to be? England had solved the question in 1797 by making the new medium bank notes. The new medium was to be paper. This plan only differed from its realization in suggesting a day-unit for an hour-unit. In 1823 he recommended 'notes representing any number of "days" 'labour or part of a day's labour' (*Reports of Meetings in Dublin*, p. 127). 'Equitable labour exchange' applies therefore to barter as well as to exchange by labour notes.

The Scheme

The First Scheme (1827–30) is often attributed to Josiah Warren, who after assisting in the disastrous communistic experiment of Owen at New Harmony (1825–7; for Owen's plan, see below) became an individualist. No account is obtainable of Warren's first experiment (*c* 1828) at Cincinnati. Warren's New Harmony experiment (1842) is thus described by Macdonald: the purchaser paid *in cash* wholesale prices plus 5 per cent for general expenses and added a promise to labour for, say, the hour during which the storekeeper attended to him; he then valued his promise in kind (or in cash?) and redeemed it accordingly (Noyes, pp. 96, 97). We do not read of the storekeeper buying with labour-notes but with cash, and in 1852 Warren said it was his rule that what was bought with cash must be sold for cash (*Equitable Commerce* by Josiah Warren, pp. 85, 91, 92, 109, etc.). The labour-notes were merely a medium for paying store servants for their trouble in kind. The only interest of the scheme is that it was a cooperative store. Warren had ulterior views no doubt, but these were to be carried out by corn-notes (see below).

In England, at that date 'cooperative society' meant a club whose members subscribed 1 s. a week or so to a 'community fund', or a fund for starting an Owenite 'village' in which producers should produce all that they wanted, and so turn communists; 'trading associations' meant cooperative stores in the modern sense based on this community fund; and 'union exchange' meant cooperative stores bought from cooperative producers. In August 1827 Dr King grafted on 'The London Co-operative Society' at 36 Red Lion Square, a 'union exchange' (*Co-operative Magazine*, ii. 421), which Lovett called 'The First London Co-operative Trading Association'. During September, Owen, then on a flying visit to England, saw Dr King and induced him to divide the community fund amongst the members each month. In announcing this change (1 December 1827), Dr King wrote of his scheme as a scheme for 'exchanging labour', which meant buying and selling at cost price, and as leading to everything Owen ever contemplated (ibid., ii. 548). It is hard to see how the Owenite ideal of economical self-sufficiency could be obtained by a group of townsmen if they gave up the plan of a permanent community fund. But there was one other possible method, alliance with other groups of cooperative producers. This method was probably present to Dr King's mind. Again on 1 October 1827 the Brighton cooperators, whose prophet was Dr King, proposed a similar exchange union with labour-notes or 'notes for value of so much labour as is brought in' (ibid., p. 511).

The example of London and Brighton spread through the kingdom, and we come to the *second scheme* (1829–34), whose *differentia* is the alliance of cooperative societies, in the modern sense, through labour-notes. On 13 January 1830, 'The British Association for promoting Co-operative Knowledge' officially proclaimed the federal idea (London, *Co-operative Magazine*, p. 28), the idea of forming what the *Quarterly Review* of November 1829 (p. 373) called 'a bazaar of cooperative shops'. Owen, who had permanently returned to England in the previous August, inspired, but did not head the new departure. On 28 April 1832, The *Crisis* advertised an 'exchange bazaar' in New Road, Marylebone, then the headquarters of the British Association, 'on an equitable time valuation', under the signatures of Dr King and Macpherson According to Lovett (*Life*, p. 47), this meant labour-notes. The commission

charged was 8 1/3 per cent = to 1d. in 1 s. Owen, his hand being thus forced by his disciples, then published his full scheme (*Crisis*, 16 June 1832), with draft labour-notes (ibid., 30 June), and rules (ibid., 30 June and 8 September). Owen's 'Institution' – as the headquarters of his 'Association of the Industrious Classes, founded 1831', were called – was at Bromley's Bazaar, Gray's Inn Road. It had been a club for ventilating unpopular religious views, but was now quickly adapted to its new purpose. Deposits began 3 September; exchanges, 17 September, and a branch office was opened 8 December in Blackfriars. The maximum deposits in the Bromley Bazaar reached 38,772 hours in one week, and after a month the branch office recorded 32,759 hours' deposits 16,621 hours' exchanges (*Crisis*, ii. 7); (Holyoake writes £ for hours). The 'institution' merged in the Blackfriars branch from January to May 1833, when it migrated to 14 charlotte Street, Fitzroy Square, whence it formed a Birmingham branch which opened on 29 July and 12 August 1833 for deposits and exchanges respectively. This 'institution' was by far the most important federal centre of the new movement, but while it invited non-members as well as members to deal with it, cooperative societies usually kept their organizations distinct from it. Owen undertook to absorb into it every trade, benefit, and cooperative society in the kingdom (*Crisis*, 14 April 1833), but a year later it abandoned industrial federalism (*Crisis*, 7 June 1834). In spite of this abandonment cooperative societies had been swept into the stream mainly by Owen. Pare (Owen's son-in-law) and Dr King turned their clubs, 'trading associations', and 'union exchanges' into 'equitable exchanges', federated throughout the length and breadth of the land, and in a year or two were nearly all engulfed (Booth says 'all but four', *Robert Owen*, p. 154, but see *Working Men Co-operators* by Acland and Jones, p. 23).

Labour Exchange Notes

'This little and apparently insignificant instrument would bring prosperity to all', – so said Owen of the *first* of the notes represented here (*Crisis*, 2 October 1832), which bears date fifteen days after the Bromley bazaar stores were opened for Exchange. The example published in Lloyd Jones's *Life of Owen*, 2nd edn, 1895, p. 240, is marked 'Birmingham Branch', has no reference to an eight hours' day (as the one given here has), but has the same pattern. 'The sun of truth' recalls the titles of two Owenite newspapers, the daily and weekly *True Sun*. The beehive commemorates one of Owen's favourite fables (*Crisis*, ii. 40). The scales of justice adorn, also, J. Warren's corn-note of 1852. The note is in form a bill of exchange, and in substance a deposit-receipt, and therefore, unlike the IOU's devised by Warren, precluded credit. It was transferable in name and fact, but not in law. The *second and third* of these notes were obviously issued by the 'London Co-operative Trading Association', and the word 'central' indicates that they too were federating. There is no trace of notes actually issued by this or any other cooperative society before April 1832. E. Nash, the secretary of Owen's central association, whose name appears on the first note, often warned people against notes issued by non-affiliated societies which worked on slightly different lines (*Crisis*, i. 143); perhaps this society was referred to. On 14 April 1833, 'extensive premises in Red Lion Square, lately the Labour Exchange and Institution for the Working Classes', were advertised for sale; these premises were apparently referred to. *The third note* marks the point where the labour-standard degenerates into or emerges from the money standard. Owen's first draft-note (*Crisis*, 20 June 1832) also stated that 'the price of labour is 6d. an hour'; but the rules explain that this superscription only meant that materials were valued thus; Warren's notes, which he suggested for general use, were as shown in Table 1.

Warren's circulating medium, which he forbade to circulate, in effect substituted corn for labour, just as this third note substitutes corn or labour as the standard of value. This note is practically a bill of exchange; only a technicality of English law prevents it being regarded as such. Labor for labor.

Labour Exchange, Table 1

Cost the limit of price	Seven hours	Not transferable	7–12 pounds	Labor for labor.
(Figure of Justice)	Due to Jacob Smith SEVEN HOURS' LABOR.			
Justice	In House Rent or SEVEN-TWELVE POUNDS of CORN.			

The Principles On Which They Were Based

As the Assignats were a paper currency based upon land, so Robert Owen proposed in 1820, and his societies tried to carry out in 1832, a currency based upon labour. His labour notes were warrants issued on the strength of an hour's labour, and entitling the holder to goods from the store of the issuing exchange 'to the value of one hour'. Articles were to be exchanged at cost price, cost being assumed to be simply the labour spent on them. For the sake of bridging over the transition from the old currency to the new, labour was valued at 6d. an hour. Thus at the labour bank in the Gothic Hall, Marylebone, those who deposited goods at the stores were paid in labour-notes according to the value of the goods as so estimated.

For several months there was every sign of success, and some hundreds of London tradesmen agreed to take the notes in payment from their customers. The prosperity of the cooperative exchanges caused the rise of spurious rival institutions which soon forfeited the public confidence, and in the course of the year brought themselves and their models to a common ruin.

In any case no permanent prosperity could have been expected. Beginning with the error of treating all value as a matter of cost and all cost as labour, the promoters of the scheme were, besides, not equal to the task of distinguishing between the hour's labour of the skilled and industrious and the hour's labour of the unskilled or the idle. To discriminate accurately by having regard to length of training and to the ease or difficulty of the labour attested by the note would have complicated a scheme of which the most vaunted merit was its simplicity. Owen himself, too, was conscious that, especially at first, the ways and even the language of ordinary business must be preserved. But his followers, with few exceptions, were without discretion, and imposition was easy. Men brought goods that were unsaleable in the ordinary market, turned them into labour-notes, and with these notes drew useful and saleable articles from the stores. If careful valuation had been made for them by a common pawnbroker, the exchange societies might, at a small expense of dignity, have purchased a longer lease of life.

The idea of a labour note was in Owen's mind as early as 1820. In the Report to the County of Lanark, 1820 (*Life*, vol. ii, 267 *seq*., he says that 'the natural standard of value is in principle human labour'), 'the average of human labour or power may be ascertained; and, as it forms the essence of all wealth, its value in every article of produce may also be ascertained, and its exchangeable value with all other values fixed accordingly, the whole to be permanent for a given period. Human labour would thus acquire its natural or intrinsic value'.

Owen continues (ibid., p. 278), 'To make labour the standard of value it is necessary to ascertain the amount of it in all articles to be bought and sold. This is in fact already accomplished, and is denoted by what in commerce is technically termed the 'prime cost', or the net value of the whole labour contained in any article of value the material contained in or consumed by the manufacture of the article forming a part of the whole labour'. 'The genuine principle of barter was to exchange the supposed prime cost of, or value of labour in, one article, against the prime cost of, or amount of labour contained in any other article. This is the only equitable principle of exchange', and it may be secured without sacrifice of modern improvements (p. 279), 'by permitting the exchange to be made through a convenient medium to represent this value'. He goes on (p. 304): 'A paper representative of the value of labour manufactured on the principle of the new

notes of the Bank of England will serve for every purpose of their [the association's] domestic commerce or exchanges and will be issued only for intrinsic value received and in store'.

It must be said that these notes cannot fairly be compared with ordinary bank notes; they were not issued for profit or on a calculation of probable demands for payment, but simply to effect the exchange of two supposed equivalents both actually existing at the time of exchange. Over-issue was impossible, for the goods might be said to go with the notes, as with bills of lading. In theory they were always convertible. If depreciation occurred, it was because of the spread of disbelief in the possibility of carrying out the conditions of the scheme, not from the nature of the case owing to an issue beyond the needs of the public.

The figure below is description of a proof of one which is preserved in a collection made by Francis Place, in four volumes of his, Owen's, and similar authors' writings ranging from 1817 to 1832 on *Labour Questions and Political Economy.*

Bibliography

Held, A. 1881. Sozialen Geschichte Englands. Leipzig.
Herbert, W. 1825. Harmony. London.
Holyoake, G.J. 1875. History of co-operation. London.
Jones, B. 1894. Co-operative production, 2 vols. Oxford: Clarendon Press.
Menger, A. 1891. *Recht auf den vollen Arbeitsertrag*, 2nd ed. Stuttgart: J.G. Cotta.
Noyes, J.H. 1870. A history of American socialism. Philadelphia.
Thompson, W. 1824. An inquiry into the principles of the distribution of wealth. London.

Labour Market Discrimination

Irene Bruegel

The facts of continued discrimination on grounds of sex and race point up some of the inadequacy of neoclassical labour market theory. The idea that pay reflects value, bar peripheral imperfections, is at odds with the experience of blacks and women in the labour market. Indeed if the newly won concepts of comparable worth and equal value now embodied in the American and British equal pay legislation were truly effective, many established pay relativities would be undermined. Neoclassical labour market theory merely adds discrimination on to its existing model but discrimination, as a structural feature of the labour market, calls up a very different approach to the analysis of labour markets.

Modern neoclassical literature on discrimination takes as its starting point Gary Becker's *Economics of Discrimination*, published in 1957, and is largely couched in the framework of human capital theory. As such it is flawed from the start by an assumption that pay, productivity and value are all three accounted for by individual attributes:- specifically education, training and experience, and that relations of power, social norms and expectations are, if anything, external issues.

A perfectly competitive economy is taken as the norm, against which discrimination by sex or race is conceptualized as an unfortunate, but peripheral, aberration based on prejudice. In orthodox economists' terms the potential for wage discrimination exists wherever *equally productive* workers receive *unequal rewards*. Such a definition of (wage) discrimination does not adequately acknowledge the social determination of discrimination and the implications of this. Discrimination by sex and race are treated as essentially parallel phenomena, amenable to the same basic analysis, even though the forces which create and sustain such discrimination may in fact be very different.

There are three interelated areas of debate in the economics of discrimination: (i) the definition of discrimination; (ii) the measurement of the scale of discrimination against any particular group; (iii) the identification of the perpetuators and beneficiaries of discrimination.

The Meaning and Measure of Discrimination

There is no consensus on how discrimination is to be defined, once one goes beyond the theoretical

definition to issues of policy. The first step is to confine the concept of discrimination to instances where the treatment of a person reflects his or her membership of a particular social group. But the recognition of a group as potentially vulnerable to discrimination is not unambiguous. It took organization and opposition to get the issue of women's pay and pattern of employment raised above the 'natural order of things'. In the same way, occupational and pay differences which are currently regarded as 'normal' – such as those between old and young or manual and non-manual labour, able-bodied and disabled, could usefully be placed within the context of discrimination.

In attempting to define or delineate discrimination three main issues arise; the relevance of the victim's choices, the discriminator's motivations and the meaning of equal productivity or value.

Unequal pay for equally valuable workers does not necessarily signal discrimination, because unequal pay may reflect other rewards. It has been argued that unequal pay which results from choosing a particular type of job, say one that fits with a feminine image or with domestic responsibilities, is not discriminatory. Polachek, for example, argues that women's lower pay results from a rational decision by them to opt for jobs with flat career profiles (Polachek 1979). Other neoclassical economists dispute whether career breaks and lesser work experience do 'explain' women's poorer pay statistically (England 1982; Beller 1982). Feminists go further, questioning whether such choices should be characterized as rational adaptations to an immutable domestic division of labour (Dex 1985) rather than evidence of the deep structuring of discrimination in a patriarchal society (Barrett 1980).

Secondly, there is an issue of motivation, of direct and indirect discrimination. Sloane (1985) argues that unequal pay arising from market processes rather than a *decision* to discriminate falls outside the economists' concept of discrimination. Such a focus on intentions, rather than outcomes, would however cut out most of the indirect discrimination that the UK and US anti-discrimination legislation at least covers in principle.

The main issue is the determination of 'equally productive workers' and 'work of equal value'. Only when people are doing the same job in exactly the same circumstances is it clear whether or not they are doing work of 'equal value'. But blacks and whites, men and women are rarely to be found working alongside one another in exactly the same circumstances, precisely because of the prevalence of discrimination. There is no evidence that black people or women are inherently less productive. So the issue becomes one of identifying differences in productivity and determining which of the processes forging such differences are to be included, or controlled for, in defining and measuring discrimination.

By and large neoclassical economists identify the social processes that render different types of labour more or less valuable to capital as operating independently, and in some sense 'prior' to the labour market. People arrive on the employer's doorstep with different attributes. The neoclassical economist then analyses the pay and conditions of different groups in relation to these attributes, with the human capital theorists' focusing particularly on education and experience, to identify whether or not and how far wage discrimination exists (Mincer and Polachek 1974; Greenhalgh 1980; McNabb and Psacharopoulas 1981, etc.). Alternatively discrimination is measured through *reverse regression* – establishing the scale of any qualifications gap at the same level of pay. The greater the number of prior attributes identified and measured and the more finely the place and type of work is differentiated between industries, corporations and individual workplaces, the lower the evident discrimination.

While there may be good policy grounds for trying to identify the distinct variety of processes which contribute to the poorer earnings of ethnic minorities and women, the model is flawed by its assumption that pay differentials divide nearly into two components: that due to differences in value (whether from non labour market discrimination, choices or natural attributes) and that due to discrimination. This assumes away the power relations which structure differential pay in the real world and the intertwined relationship between pre and post labour market discrimination.

In practice differences in motivation, training, domestic duties, location etc. of second-class workers reflect actual and anticipated labour market discrimination. Women and black people work in distinct industries and occupations to some degree at least because of actual and anticipated discrimination; they may also train less and have lower motivation because potential discrimination reduces the returns to them. Measuring discrimination by differences in pay between races and sexes within set occupations and industries, then ignores both these issues. It also assumes that pay differences *between* industries and occupations arise only from differences in the productivity of labour. But if employment discrimination is prevalent such an assumption is unlikely to hold. Nor does the human capital evidence – such that it is – that pay varies with experience and education negate this point. For the return from each extra year's employment may have more to do with the typical pattern of nonmanual white men's employment than with any increments to productivity.

The problem of the 'residual' view of discrimination is illustrated by Chiswick's analysis of the position of American Jews (Chiswick 1983). Chiswick, using a standard human capital model of the type used extensively to identify the level of discrimination against blacks and women, finds that, after standardization, Jewish men earn 16 per cent more than average. He sensibly avoids a conclusion of discrimination in favour of Jews, but, short of evoking a Jewish spirit or 'X efficiency', is left without an explanation. Some refinement of the data might lower the unexplained residual below 16 per cent, but in view of the huge range of estimates of rate and sex discrimination provided by human capital models (Lloyd and Niemi 1979; Chiplin and Sloane 1982), the theorization must be open to question.

Who Benefits?

The neoclassical *explanation* of discrimination, how it arises and who benefits, is also problematic. Becker's original model (Becker 1957) takes two forms; the first derived from international trade theory and the second from utility maximizing preference theory. In both versions discrimination is posited as a cost to the discriminator; an irrational decision within an essentially rational market. Although these are made to look like results of analysis, they really stem from the specification of the models. If, as Becker assumes, whites indulge their 'taste' for discrimination discrimination by restricting the export of capital to black 'society' (i.e. by not employing blacks) then given Becker's assumptions, standard trade theory will give the result that white 'society' will lose as a whole, even though white labour and black capital may benefit from such restrictions (Madden 1973).

The relevance of such a model to an economy where blacks live and work amongst whitesbut in inferior jobs-is clearly open to question. The application of this model to sex discrimination, where men and women live jointly in households is still more questionable. Furthermore once Becker's basic assumption – that whites/men own all available capital – is explored, it becomes clear that blacks and women are forced to accept the terms of white/male society. Including such power relations gives the result that whites/males benefit from discrimination (Thurow 1976).

The microeconomic foundations of Becker's model are also suspect. Discrimination is said to arise from a 'taste' for discrimination (a distaste for employing blacks and women) on the part of employers, though the basis for such tastes and what might cause them to alter is never explored.

Using Becker's model with his assumptions again produces the result that discriminators lose out; employers who irrationally refuse to employ blacks or women face higher costs and lower profits. But what also follows is that discriminators would be driven out of business in a perfectly competitive market by employers with a lesser or different taste for discrimination. Thus the continued existence of discriminatory practices throws the model into question.

Developments in the model of individual-employer-based discrimination allow for employers to benefit from their actions. So-called statistical models of discrimination

also allow it to be rational profit maximizing behaviour (Aigner and Cain 1977). Discrimination arises because employers do not know the true value of 'minority' labour power. Since information costs money, it is rational to extrapolate the costs of employing a given individual from knowledge (or assumptions) about the characteristics of their group. There remains a problem, however, in explaining persistent discrimination since once one firm recognizes the value of minority labour, all others in competition would be forced to follow suit.

Madden (1973) shows how a monopsonist can exploit womens' lower elasticity of supply and thus benefit from discrimination. The persistence of sex discrimination can thus be explained as a result of women's lower mobility and lesser unionization. The monopsony model does hint at the importance of relations of power and differential power in explaining persistent discrimination for it implies that discriminatory wages arise from the inferior market power of discriminated groups. But it is neither a satisfactory model of race discrimination nor of sex discrimination outside the context of monopsony.

An adequate theory of discrimination would be based in a model of the labour market that encompasses relations of power, not just between employers and workers or the state and workers, but also between groups of workers who for whatever reasons, differ in their immediate interests. Historical analysis (Cockburn 1986; Hartman 1976; Humphries 1977) has helped to establish how differences of interest between male and female labour are created and sustained. For a variety of reasons male workers and white workers have identified their interests with the exclusion of competing groups of 'cheap' labour. That exclusionary discrimination has differentiated the labour market by sex and race. The fracturing of the working class in this way shifts the balance of class power to employers, so whatever the immediate costs of excluding cheap labour, discriminatory divisions have been pursued by white capital.

The *crowding* of 'second class' labour into a small set of specific jobs (Bergman 1971) and the creation of a 'segmented labour market' means that women and black people are rendered cheaper labour power. This is achieved even in the absence of overt discriminatory practices through the cultural determination of 'suitable jobs'. This does not mean that powerful anti-discrimination legislation and enforcement provision can have no effect on labour market outcomes. Were they to be put into effect, they could. However, a narrow-minded focus on the 'labour market discrimination' identified through neoclassical theory is of limited relevance since it skirts over the entrenched determination of inequalities.

See Also

- ▶ Gender
- ▶ Inequality Between the Sexes
- ▶ Occupational Segregation
- ▶ Segmented Labour Markets

Bibliography

Becker, G. 1957. *The economics of discrimination*. Chicago: University of Chicago Press.

Beller, A. 1982. Occupational segregation by sex: determinants and changes. *Journal of Human Resources* 17(3): 371–392.

Bergman, B. 1971. The effect on white incomes of discrimination in employment. *Journal of Political Economy* 79(2): 294–313.

Chiswick, B.R. 1983. The earnings and human capital of American Jews. *Journal of Human Resources* 18(3): 313–336.

Cockburn, C. 1986. *The machinery of domination*. London: Pluto Press.

Dex, S. 1985. *The sexual division of work*. Brighton: Wheatsheaf Books.

England, P. 1982. The failure of human capital theory to explain occupational sex segregation. *Journal of Human Resources* 17(3): 358–370.

Greenhalgh, C. 1980. Male–female wage differentials in Great Britain: Is marriage an equal opportunity? *Economic Journal* 90: 751–775.

Hartman, H. 1976. Capitalism, patriarchy and job segregation by sex. In *Women and the workplace*, ed. R. Blaxall and B. Reagan. Chicago: University of Chicago Press.

Humphries, J. 1977. Class struggle and the persistence of working class families. *Cambridge Journal of Economics* 1: 241–258.

Lloyd, C., and B. Niemi. 1979. *The economics of sex differentials*. New York: Columbia University Press.

McNabb, R., and G. Psacharopoulas. 1981. Racial earnings differentials in the UK. *Oxford Economic Papers* 33: 413–425.
Madden, J.F. 1973. *The economics of sex discrimination*. Lexington: Lexington Books.
Mincer, J., and S. Polachek. 1974. Family investments in human capital: Earnings of women. *Journal of Political Economy* 82(Pt. 2): 118–134.
Polachek, S. 1979. Occupational segregation among women: Theory, evidence and a prognosis. In *Women in the labor market*, ed. C.B. Lloyd, E.S. Andrews, and C.L. Gilroy, 137–157. New York: Columbia University Press.
Sloane, P. 1985. Discrimination in the labour market. In *Labour economics*, ed. D. Carline. Harlow: Longman.
Thurow, L.C. 1976. *Generating inequality*. New York: Basic Books.

Labour Market Institutions

Richard B. Freeman

Abstract

Labour market institutions – unions, collective bargaining, government regulations – that help determine wages and working conditions differ greatly across countries. Advanced European countries rely extensively on institutions while the United States relies more on market forces. Labour institutions reduce the dispersion of pay and income inequality but have problematic effects on other aggregate economic outcomes, such as unemployment. The weak or inconclusive link between institutions and outcomes beyond wage dispersion could reflect different institutional effects under different economic conditions; efficient bargaining that balances the adverse and positive effects of institutions on those outcomes, or weaknesses in data and modelling.

Keywords

Coase Theorem; Collective bargaining; Employment at will; Employment protection; Gini coefficient; Health insurance; Labour market institutions; Layoffs; Minimum wages; Rent seeking; Reservation wages; Trade union density; Trade unions; Unemployment; Unemployment insurance; Wage differentials; Wage dispersion; Wage drift

JEL Classification

J41

Labour market institutions – the organizations and procedures through which workers, firms, and the government affect wages, employment and working conditions – vary widely across countries and among firms and industries within a country. In some countries or settings within a country, trade unions, employer federations, personnel and human resource departments of firms and various forms of collective bargaining, or government regulations greatly affect how firms and workers interact at work places and help determine the hours, wages, occupational health and safety conditions, rules for promotion, and other conditions of work life. In other settings and countries these institutions have little impact. In those situations the market rules.

Once a minor tributary of economic analysis, the study of labour market institutions moved to the mainstream of discourse in the 1990s and 2000s as economists focused on differences in labour institutions as a possible cause of the varying economic performances among countries that had roughly similar macroeconomic policies. The Organisation for Economic Co-operation and Development's influential 1994 Jobs Study (OECD OECD 1994a, b) spurred research in advanced countries with its claim that many institutional interventions in the labour market reduced employment and that OECD countries should deregulate labour markets and weaken welfare state protections to achieve full employment. Ensuing analyses questioned the evidentiary basis for this diagnosis, producing a wide-ranging debate about how labour institutions affect advanced market economies. In developing countries, the analogous claim has been that institutionally determined wages and rules of work in the formal sector of economies reduce job creation in that sector and thus contribute to a dual labour market that harms economic growth and worsens

the distribution of income. This also has generated considerable debate, pitting analysts who see institutions largely as creating distortions in competitive markets against those who see them as mechanisms for resolving market failures and shifting income distribution to workers.

Institutional Differences

The starting fact for the debate is the wide variation of institutional arrangements in both advanced countries and developing countries. Table 1 summarizes the institutional architecture of the labour market in the United States and in advanced European countries – defined as European Union (EU) countries exclusive of the United Kingdom and Ireland, whose institutions are often closer to those of the United States than to the rest of the advanced Europe (Freeman et al. 2007) and inclusive of Norway and Switzerland, which are outside the European Union. The exhibit shows that the percentage of workers in unions is three times greater in the advanced European countries than in the United States. It notes a large difference in the organization of firms into employer associations. In advanced Europe many firms join employer associations that negotiate with unions, whereas in the United States employers negotiate separately with unions or with individual employees in the absence of collective bargaining. In addition, many advanced European governments extend the terms of a contract between an employer federation or major employer to all firms and workers in a sector, including those who were not party to the agreement, on the grounds that collective bargaining should produce a single wage just as supply and demand should produce a single wage in a competitive labour market. As a result of mandatory extension of contracts, the rate of collective bargaining coverage in advanced Europe (80% in the table) exceeds the rate of unionization (38% in the table); whereas the rates of union density and collective bargaining coverage are about the same in the United States. As a result, the gap in coverage between the United States and advanced Europe exceeds that in union density. The effect of mandatory extension on wage setting is most dramatic for France, where approximately 90% of workers are covered by

Labour Market Institutions, Table 1 Labour market institutions in the market-driven United States versus institution-driven advanced Europe

	USA	Advanced Europe*
Union density, 2003	12 %	38 %
Extent of employer federation	Negligible	Substantial, bargain regularly
Percentage of workers covered by collective bargaining, 2000	14 %	80 %
Extension of collective contracts	none	Widespread by law
Employment protection legislation (higher values imply more protection, from 0 to 4)	0.7	2.7
Works councils	None	Mandated
Social dialogue	None	Widespread
Ratio of unemployment insurance to past wage, 2004	54	69
Months of unemployment insurance coverage, 2004	6 months	22 months
Social expenditures as share of national income, 2003	18.7 %	28.9 %
Rating of labour market in market orientation, 2003 Fraser (1 = most market oriented), 103 countries	10	76
Rating of labour market in market orientation, 2003 Global Labor	6	26
Survey (1 = most market oriented), 33 countries		

Source: Union density from Visser (2006), OECD (2004, Table 3.3; 2004, Table 2.A2.4, version 2; 2006a, Table 3.2; 2006b, Figure GE1.2, p 41), Gwartney et al. (2005), and Freeman and Chor (2005)

*Excluding the United Kingdom and Ireland. Italy, France, Spain, Greece, Sweden, Portugal, Germany, Belgium, Austria, Denmark, Finland, Norway, Netherlands, Switzerland

collective bargaining even though union density is six per cent or so – the lowest among advanced countries.

At the enterprise level, all countries in the European Union require that firms above a specified size introduce a works council of democratically elected employee representatives and that the firm consult with the council on key decisions that affect workers. In Germany firms must reach agreement with the council on some issues or go to arbitration to resolve disagreements. By contrast, the United States outlaws non-union employee organizations at the workplace for fear that they will become company-dominated barriers to independent unions. Many US firms set up employee involvement committees to deal with issues regarding workplace productivity, but these committees cannot legally represent workers' interests to management. Going beyond enterprises, the EU relies extensively on *social dialogue* among employer federations, unions, and in many cases, governments to determine labour and other economic policies. Social dialogue produced Ireland's 1987 Solidarity Wage Agreement in which the government agreed to lower taxes on workers, unions agreed to moderate wage demands, and employers agreed to seek to increase employment. The ensuing economic boom in Ireland suggested to some that the social pact contributed positively to Irish economic performance.

There are also large country differences in hiring and firing practices. Firms in the United States operate largely by employment at will, which means that the firm can replace workers for any business or other (non-discriminatory) reason. By contrast, many EU countries have employment protection legislation that requires firms to give substantial severance pay to laid off workers and to negotiate 'social contracts' with works councils to help laid off workers obtain training and new employment. In addition, European welfare states pay higher unemployment insurance in relation to wages for longer periods of time than does the United States, and provide national health insurance that US firms and workers must fund for themselves. These policies produce higher government social expenditures as a share of national income in the EU countries than in the United States, and commensurately higher taxes as a share of national income to pay for the benefits.

Taking these and related differences in the labour market together, analysts have created aggregate thermometer style indices of the institutional versus market orientation of country labour markets, in which higher scores reflect greater reliance on markets than on institutions. The Fraser Institute – a conservative think tank that produces an index of economic freedom based on metrics for 'personal choice, voluntary exchange, freedom to compete, and protection of person and property' (Gwartney et al. 2005, p. 5) – codes countries that have extensive legal protection of labour and high levels of collective bargaining as having less economic freedom than those without these institutions. The Global Labor Survey has created a comparable index by asking union leaders, labour relations professors and other experts to report on the *actual situation* of labour in their country (Chor and Freeman 2005). The difference in ideological persuasion between the Fraser Institute and most respondents to the Global Labor Survey notwithstanding, the two indices tell a similar story about cross-country differences. They give the United States and the other English-speaking advanced countries higher scores in using markets than European Union economies, and give the Scandinavian countries, which rely extensively on collective bargaining to determine pay and working conditions, particularly low scores in reliance on markets. While analyses of labour institutions in developing countries are less plentiful, the Fraser Economic Freedom Index and Global Labor Survey show a similar wide variation in the institutional framework for those countries. Botero et al. (2004) provide additional information on labour institutions across countries in terms of their labour laws. The indices of labour laws measure *de jure* labour institutions, whose impact on the labour market depends on the extent to which countries enforce their laws.

Institutions and Outcomes

To see how institutions affect economic outcomes, analysts compare the economic outcomes

for firms and workers *within countries* whose pay and work conditions are set by unions or regulations with the outcomes of firms and workers whose pay and conditions are set by market forces; compare the outcomes when institutional rules change (for instance, through an increase in minimum wages); and when the workers or firms move from market determination of wages and conditions of work to having an institution determine wages and conditions, or vice versa (for instance from moving from union to non-union status or non-union status to union status). To analyse how differences in institutions affect outcomes across countries, analysts contrast labour market outcomes *between countries* that rely more on institutions and those that rely more on markets; and compare outcomes before and after a country changes its institutions with outcomes in countries that maintain their institutions over the same time period. The goal is to use the experiences of countries that do not change institutions as a counterfactual to predict what might have happened to countries that change institutions, and, conversely, to use the experience of the country that changed institutions to predict what might have happened in countries with stable institutions if they were to change.

Constructing a counterfactual to assess the impacts of institutions is difficult. One difficulty is that changes in institutional arrangements can affect the behaviour and outcomes for the group that is not covered by the changes as well as the covered group. A decline in union density, for example, might lower the wages of union and non-union workers equally, so that the differential between them was constant, which an analyst could misinterpret as implying no change in the wages of union workers. Another reason is that persons involved with institutions learn from past experiences, so that they may respond differently in the future to a given change in conditions than they might have done in the past. British unions made different decisions in the 1990s from those they made in the 1970s, in part because of their experiences in the earlier period. Finally, to the extent that one institutional rule affects another, a counterfactual analysis of a change in a single institution can be misleading if it does not allow for how the change interacts with other regulations and rules. When Spain enacted a law permitting firms to hire workers on temporary contracts, there was a huge increase in the proportion of workers hired under those contracts. When Germany enacted such a law, firms continue to hire apprentices for permanent jobs.

Difficulties of developing a valid counterfactual notwithstanding, virtually all analyses find that labour institutions reduce the dispersion of hourly earnings and the inequality of income (which depends on hours worked, and streams of income outside of work in addition to hourly pay) compared to market-pay setting. Studies that compare the distribution of earnings and incomes across countries find, for example, that the pay of persons in the 90th percentile of wages and salaries in relation to the pay of persons in the 10th percentile is lower in the advanced European countries that rely more on collective bargaining than in the market-driven United States and other English-speaking countries; and that the Gini coefficient of inequality for total income is also markedly lower in countries where labour institutions dominate wage-setting (Table 2). The US has the largest 90/10 earnings ratio of wages and the largest Gini coefficient for total income among advanced countries. By contrast, the Nordic countries, where collective bargaining sets wages for the vast majority of workers, have the lowest dispersion of pay and low Gini coefficients. Other advanced European countries and Japan also have relatively low pay dispersion and Gini coefficients. Centralized collective bargaining arrangements are sufficiently effective to narrow pay gaps even though most centralized agreements allow for 'wage drift' – higher or lower wages for some firms and workers than the negotiated central agreement due to variations in local market conditions.

Looking at earnings when country institutions change, increased reliance on institutions narrows the distribution of earnings while increased reliance on market- wage setting widens the distribution. Declines in collective bargaining coverage in the United States, Canada, United Kingdom and New Zealand contributed to greater inequality in those countries. Similarly, the decline in the real

Labour Market Institutions, Table 2 90/10 Wage differentials and Gini coefficients for advanced countries, circa 2000

	Dispersion	Gini
US	4.59	40.8
Other English-speaking	3.46	35.2
Advanced Europe	3.10	32.2
Japan	2.99	24.9
Scandinavia	2.18	25.6

Source: 90/10 ratios averaged from data from OECD (2004, Table 3.2), where the data are from 1995 to 1999 with figures from Austria, Belgium, Denmark, Portugal are for 1990–1994; data for Spain and Greece from Martins and Pereira (2004, Table 1). Gini coefficients from United Nations, Human Development Report (2005, Table 15) Other English-speaking countries are: the United Kingdom, New Zealand, Canada, Ireland and Australia. Advanced Europe countries are: Belgium, Netherlands, Italy, Switzerland, France, Austria, Germany, Spain, Portugal, Greece; Scandinavia are: Norway, Finland, Sweden and Denmark

value of the US minimum wage added to inequality, while the introduction of the minimum wage in the United Kingdom limited the rise of inequality in that country. The breakdown of centralized negotiations between the major union federation and major employer association in Sweden raised inequality modestly in that country. But perhaps the most compelling evidence comes from the rise and fall of Italy's Scala Mobile mode of pay setting. The Scala Mobile was a national agreement that gave larger percentage increases in pay to low-wage workers than to high-wage workers. When the Scala Mobile determined wages, the dispersion of earnings in Italy fell sharply – towards Scandinavian levels. When Italy abandoned this mode of pay setting, in part because the distribution of pay seemed to have narrowed wage differentials beyond what made economic sense, the dispersion of earnings increased (Erickson and Iquino 1995; Manacorda 2004).

Studies within countries that contrast the inequality of pay among workers whose pay is set by institutions and those whose pay is set by markets also find that institutions are associated with lower dispersion of pay. Dispersion is less among unionized workers than among otherwise comparable non-union workers and less among government employees than among private sector employees whose pay is market-determined. Moreover, although the wage differential between union and non-union workers raises inequality between organized and non-organized workers, the net effect of unions on earnings is to reduce inequality. The overall distribution of earnings is dominated by the compression of wages within the union sector and by difference the reduced earnings between management and other high-paid nonunion workers and union workers within firms. Consistent with this, studies that contrast the inequality of pay among workers who shift from non-union jobs to union jobs or the converse find that dispersion among a group of job changers falls when workers enter the union sector and rises when they leave the unionized setting (Freeman 1984).

Is the institution-induced reduction in the dispersion of pay good or bad for the economy? To the extent that real world labour markets perform largely as ideal competitive markets, the reduced dispersion of pay distorts economic decisions on both the supply and demand sides of the market. By contrast, to the extent that real- world labour markets fall short of the competitive ideal, institutions can improve the efficiency of markets. There are plausible arguments and evidentiary support for both interpretations of what institutions do.

The Arguments

The claim that *labour institutions adversely affect economic performance* begins with the assumption that in the absence of institutional interventions, real labour markets produce wage, employment, and working conditions that approach those of an ideal competitive labour market. In this case, institutions can only distort

incentives and reduce the efficient allocation of resources. For instance, union-induced wages above the market rate induce unionized firms to reduce employment, which reallocates labour to lower paid less productive activities. The following statement from the World Bank expresses the view that institutions distort the demand for labour in developing countries and slow down the shift of labour from agriculture and informal sector work to more highly productive and better-paid formal sector jobs:

> Labor market policies – minimum wages, job security regulations, and social security – are usually intended to raise welfare or reduce exploitation. But they actually work to raise the cost of labor in the formal sector and reduce labor demand . . . increase the supply of labor to the rural and urban informal sectors, and thus depress labor incomes where most of the poor are found. (World Bank 1990, p. 63)

The arguments against labour institutions in advanced countries are similar. On the demand side, institutionally driven increases in wages for the low-paid raise their cost to employers, which lowers their employment, and distort the allocation of the workforce among sectors, squeezing in particular low wage service industries. On the supply side, institutionally driven reductions in earnings inequality reduce pecuniary incentives to make efficient economic decisions. All else the same, reductions in the earnings premium paid to more-skilled workers will reduce investments in skills. And high unemployment insurance benefits will induce laid off workers to raise the reservation wage at which they will accept a new job and to search less intensely for jobs, producing longer spells of joblessness and higher rates of unemployment. In addition, the reduction in job search will lessen supply side pressures towards modest wage settlements that help job creation.

The magnitude of the distortions depends on the responsiveness of decision-makers to the institutionally determined incentives. In the standard 'welfare triangle' analysis, the economic loss from raising a wage above the market rate depends on the magnitude of the wage change and the elasticity of demand, which determines the magnitude of the distortion in the allocation of labour. (The formula for a welfare loss is ½ (change in wages) × (change in employment), where the change in employment is the elasticity of demand times the change in wages.) The higher the elasticity of demand, the greater will be the welfare loss from wages above the market rate. Similarly, on the supply side, the higher the elasticity of supply to the returns to skills, the greater will be the welfare loss from decisions to forgo investments in skill due to the compression of wages, and the higher the elasticity of supply to unemployment benefits, the greater will be the welfare loss, due to the decision to search less intensely for a new job due to unemployment insurance.

Finally, institutional determination of labour outcomes can impose two additional costs on the economy. The first is the political lobbying and related resources that labour and management spend to affect labour regulations and the rules governing union and employer interactions. These are sometimes pejoratively labelled as the costs of rent seeking, though if institutions help to solve economic problems they could just as well be called the costs of problem-solving. The second are the resources involved in implementing institutional arrangements. These range from establishment of union and employer federations, time spent in negotiations and dialogue at the workplace and at national levels. Discussion reduces the speed of decision-making, so that institutionally driven systems are likely to respond more slowly to economic changes than market-driven systems.

On the other side of the debate, the argument that *labour institutions improve economic performance* begins with the belief that real labour markets fall short of competitive equilibrium. Analysts view the high dispersion of pay for workers with observationally equivalent skills as reflecting the failure of the market to establish a single price of labour for similar workers. If this is a correct reading of the data, institutionally determined reductions in dispersion could create outcomes closer to the competitive ideal just as institutionally determined increases in wages can

induce firms that are monopsonies to raise employment to competitive levels. Looking at the dynamics of wage changes, in an ideal competitive system, improvements in productivity in a given sector are supposed to show up in lower prices to consumers, not in higher wages (Salter 1960; Council of Economic Advisors 1962); while changes in the prices of products due to changes in demand are supposed to induce firms to change output and employment but not to change wages. The reason wages are not expected to respond to these shocks is that the competitive model posits that firms face a perfectly elastic supply of labour at the market wage rate. In fact, changes in wages are highly related to changes in productivity and prices among industries in the United States but not in the Nordic and other countries where institutions determine wages (Holmlund and Zetterberg 1991; Teulings and Hartog 2002). At the national level, some analysts argue that the union and employer federations that negotiate national wage agreements adjust wages more rapidly to macroeconomic developments such as balance of payments or inflation than local labour markets that respond to the macroeconomy less directly.

Finally, inside firms, labour institutions can facilitate the flow of information from workers to management and from management to workers. Workers are more likely to provide information to management when they can influence how management uses the information. Regulations or union pressure that force management to open its books to workers gives them or their representative access to the same information that guides management. Increasing the flow of information and communication can in turn lead management and workers to make better decisions. Workers will be more likely to give wage concessions when the firm is truly in crisis and avoid being snookered when the firm cries 'wolf' while continuing to earn profits (Freeman and Lazear 1995). In addition, workers who have an institutional voice for dealing with problems are less likely to quit their employer and more likely to invest in firm-specific skills and seek to resolve problems by bringing them to the attention of management.

Evidence

The *OECD Jobs Study* contains two volumes of research and references to research that buttressed its claim that labour institutions explained some of the job market problems of OECD countries. Since the *Jobs Study* many other analysts have examined the link between those institutions and outcomes, generally using cross-country time series data that the OECD provides. Each year the OECD reviews the latest findings on particular issues regarding the impact of labour institutions in its *Employment Outlook*. As economists inside and outside the OECD have critically examined the data and models that link outcomes to institutions, they have moved to a more cautious stance about the evidentiary support for the *Jobs Study* conclusions. Assessing the time series models that the OECD and others used in their analyses, Baker et al. (2005) found that the estimated coefficients on labour institutions were not robust to changes in specification. They found that models that covered more years, additional countries or used different measures of the institutions than the early studies 'provide little support for those who advocate comprehensive deregulation of OECD labour markets' (2005, p. 106). Baker et al. conclude that there is a 'yawning gap between the confidence with which the case for labour market deregulation has been asserted and the evidence that the regulating institutions are the culprits' (2005, p. 198). Assessing results in the mid-2000s, Howell et al. (2007) and Baccaro and Rei (2005) come to a similar conclusion.

For its part, the OECD has recognized that the evidence is more equivocal than first claimed. The 2004 *OECD Employment Outlook* noted that 'the evidence of the role played by employment protection legislation (EPL) on aggregate employment and unemployment rates remains mixed' (2004, p. 81). It argued for 'the *plausibility* (my italics) of the Jobs Strategy diagnosis that excessively high aggregate wages and/or wage compression have been impediments' to jobs, while admitting that 'this evidence is somewhat fragile'. With respect to unionism, it summarized research as showing the effect of collective

bargaining 'to be contingent upon other institutional and policy factors that need to be clarified to provide robust policy advice' (2004, p. 165). In a similar vein, the IMF (2003) reported that 'Institutions ... hardly account for the growing trend observed in most European countries and the dramatic fall in U.S. unemployment in the 1990s.' German unemployment, for example, rose by about six percentage points in the 1990s while US unemployment fell, even though labour institutions were broadly unchanged in both countries. But the IMF still concluded that the route to full employment rested with deregulating labour markets. The strong priors and commitment to the case that institutions are the problem overrode the actual evidence.

The 2006 *OECD Employment Outlook* went a step further in assessing the impact of institutions on outcomes. It highlighted that countries with low unemployment had very different modes of wage-setting, ranging from some smaller European countries that relied on collective bargaining to the more market- determined United States and United Kingdom (2006a, Table 6.3). If different institutions can reach similar market outcomes, there may be no 'peak' form of labour market institutions to which each country should strive (Freeman 2002). But this does not resolve the debate over the impact of institutions. In a study that took account of criticisms of the non-robust findings of earlier cross-country time series data, Bassanini and Duval (2006) found that changes in tax and labour policies explain about half the 1982–2003 changes in unemployment among countries, with changes in tax policies playing a particularly large role.

The potential effect of employment protection legislation on unemployment has attracted considerable attention. Countries pass these laws to reduce layoffs and raise job security for existing workers. But the laws make it more expensive to hire workers since firms must factor in the greater expense of laying them off if business dictates reductions of output. The net effect of employment protection laws on aggregate employment thus depends on the degree to which they reduce layoffs compared to the degree to which they reduce hires. An alternative perspective predicts that on net the employment protection laws should have little or no impact on aggregate employment or unemployment. If employers and unions bargain efficiently, then the Coase Theorem predicts that they should bargain so that the firm makes the efficient layoff regardless of the employment protection law. What differs is the division of the profits from the efficient choice. With employment protection the firm pays some of the profit from a layoff to the worker to get the worker to leave. With employment at will, the firm gets all the profit from the decision. Studies of unemployment and employment between countries with greater or lesser employment protection are broadly consistent with this view. They show that the regulations have little effect on the overall rate of unemployment but shift unemployment from older workers to younger job-seekers (OECD 2004). In developing countries as well, job security regulations appear to shift employment from the unskilled youth to the skilled and older workers protected by the legislation. (Montenegro and Pages 2003).

In summary, there is no clear consensus from the empirical analyses that labour institutions have adverse or positive effects on aggregate economic outcomes beyond their distributional effects on earnings or employment.

Alternative Interpretations

There are three possible interpretations of the empirical evidence that institutions reduce the dispersion of earnings and income but do not have clear or easily identified effects on other aggregate economic outcomes.

The first interpretation is that extant measures of institutions and models of their impact are too crude to pin down the hypothesized effects on other outcomes. Better cross-section time series data on countries and more sophisticated statistical modelling might produce statistically significant impacts of institutions on outcomes beyond dispersion of pay. Most economists believe that disaggregated data that cover thousands of

observations on individuals or firms has a greater likelihood of pinning down responses of individuals and firms to changes in labour policies and institutions than further analysis of short time series across countries. But these analyses are insufficient in themselves to capture what might happen when a country changes its institutions. What might better illuminate the impacts of institutions at the national level would be to combine estimated response parameters from microeconomic studies with artificial agent models that simulate labour markets under different institution.

The second interpretation is that the effects of institutions vary over time as the economic environment changes. Given that the labour market institutions in the United States and advanced Europe were largely unchanged between the 1960s and 1990s, the only way for institutional factors to explain lower European unemployment in the former period and higher European unemployment in the latter period would be that the impact of the institutions changed over time (Blanchard and Wolfers 2000; Lundquist and Sargent 1998; OECD 2006a, b). Perhaps EU institutions were well suited to produce low unemployment in the economic conditions of the 1960s–1980s while US institutions were better suited to produce low unemployment in the globalized digital economy of the 1990s and 2000s. This interpretation is appealing. But it is difficult to test since it makes great demands on data. Allowing institutions to affect outcomes differently in different time periods reduces the number of observations with which to test the hypothesized impact and risks creating epicycles of interactions to account for observed patterns.

The third interpretation is that in fact labour institutions have first-order effects on income distribution but only modest second-order effects on other outcomes. Perhaps the hypothesized adverse effects that institutions can have on economic efficiency are balanced by their hypothesized positive effects, giving a net effect around zero. This is consistent with efficient bargaining theory, in which parties strive to reach efficient outcomes but battle over distribution. This interpretation is appealing. But there are enough situations in which unions, firms and governments do not reach efficient solutions to raise questions about it. As Sir John Hicks pointed out in *The Theory of Wages* (1934), efficient bargaining implies that strikes, which in most cases harm workers and firms, should vary randomly across industries, regions, firms and time as a result of random errors of judgment or communication. In fact strikes occur frequently in some sectors (for instance coal mining) and not in others, in some firms but not in others, and vary over the business cycle in ways that conflict with the efficient bargaining model.

In conclusion, we need to learn much more about how labour institutions affect the economy and how they operate for us to resolve the debate over whether institutions are part of the problem facing economies or part of the solution, or, more likely, which institutions and issues fall more into the former category and which fall into the latter category under particular economic conditions.

See Also

- Labour Market Search
- Labour Supply
- Minimum Wages
- Unemployment
- Unemployment and Hours of Work, Cross Country Differences
- Unemployment Insurance

Bibliography

Baccaro, L. and D. Rei. 2005. *Institutional determinants of unemployment in OECD countries: A time series cross-section analysis (1960–98)*. Discussion Paper No. DP/160/2005. Geneva: International Institute for Labor Studies.

Baker, D., A. Glyn, D. Howell, and J. Schmitt. 2005. Labor 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 policies and institutions*. Working Paper No. 486. Paris: Economic Department, OECD.

Blanchard, O., and J. Wolfers. 2000. Shocks and institutions and the rise of European unemployment: The aggregate evidence. *Economic Journal* 110: 1–33.

Botero, J., S. Djankov, R. La Porta, and F. López de Silanes. 2004. The regulation of labor. *Quarterly Journal of Economics* 119: 1339–1382.

Chor, D. and R. Freeman. 2005. *The 2004 global labor survey: Workplace institutions and practices around the world*. Working Paper No. 11,598. Cambridge, MA: NBER.

Council of Economic Advisors; 1962. *Economic report of the president*. Washington, DC: GPO.

Erickson, C., and A.C. Iquino. 1995. Wage differentials in Italy: Market forces, institutions and inflation. In *Differences and changes in Wage structure*, ed. R.B. Freeman and L.F. Katz. Chicago: University of Chicago Press for NBER.

Freeman, R.B. 1984. Longitudinal analyses of the effects of trade unions. *Journal of Labor Economics* 2: 1–26.

Freeman, R.B. 1993. Labor market institutions and policies: Help or hindrance to economic development? *Proceedings of the world bank annual conference on development economics 1992*. Washington, DC: World Bank.

Freeman, R.B. 2002. Single peaked vs. diversified capitalism: The relation between economic institutions and outcomes. In *Advances in macroeconomic theory*, ed. J. Drèze. London: Palgrave.

Freeman, R.B. 2005. Labor market institutions without blinders: The debate over flexibility and labour market performance. *International Economic Journal* 19: 129–145.

Freeman, R.B., and E. Lazear. 1995. An economic analysis of works councils. In *Works councils: Consultation, representation, and cooperation in industrial relations*, ed. J. Rogers and W. Streeck. Chicago: University of Chicago Press for NBER.

Freeman, R.B., P. Boxall, and P. Haynes. 2007. *What workers say: Employee voice in the Anglo-American world*. Ithaca: Cornell University Press.

Gwartney, J., R.A. Lawson, and E. Gartzke. 2005. *Economic freedom of the world: 2005 annual report*. Vancouver: Fraser Institute.

Hicks, J. 1934. *The theory of wages*. London: Macmillan.

Holmlund, B., and J. Zetterberg. 1991. Insider effects in wage determination: Evidence from five countries. *European Economic Review* 35: 1009–1034.

Howell, D.R., D. Baker, A. Glyn, and J. Schmitt. 2007. Are protective labor market institutions at the root of unemployment? A critical review of the evidence. *Capitalism and society* 2, Article 1. Abstract online. Available at http://www.bepress.com/cas/vol2/iss1/art1. Accessed 12 June 2007.

IMF (International Monetary Fund). 1999. Chronic unemployment in the Euro area: Causes and cures. In *World economic outlook*, May. Washington, DC: IMF.

IMF. 2003. Unemployment and labor market institutions: Why reforms pay off. In *World economic outlook*, April. Washington, DC: IMF.

Kruse, D. 1993. *Profit-sharing: Does it make a difference?* Kalamazoo: W.E. Upjohn Institute for Employment Research.

Layard, R., S. Nickell, and R. Jackman. 1994. *The unemployment crisis*. Oxford: Oxford University Press.

Lundqvist, L., and T.J. Sargent. 1998. The European unemployment dilemma. *Journal of Political Economy* 106: 514–550.

Manacorda, M. 2004. Can the Scala Mobile explain the fall and rise of earnings inequality in Italy? A semi-parametric analysis, 1977–1993. *Journal of Labor Economics* 22: 585–613.

Martins, P.S., and P.T. Pereira. 2004. Does education reduce wage inequality? Quantile regression evidence from 16 countries. *Labour Economics* 11: 355–371.

Milner, H., and E. Wadensjö. 2001. *Gösta Rehn, the Swedish model and labour market policies: International and national perspectives*. Aldershot: Ashgate.

Montenegro, C.E. and C. Pages. 2003. *Who benefits from labor market regulations? Chile 1960–1998*. Working Paper No. 9850 Cambridge, MA: NBER.

Nickell, S.J., and B. Bell. 1996. Changes in the distribution of wages and unemployment in the OECD countries. *American Economic Review* 86: 302–308.

Nickell, S., L. Nunziata, and W. Ochel. 2005. Unemployment in the OECD since the 1960s: What do we know? *Economic Journal* 115: 1–27.

OECD (Organisation for Economic Co-operation and Development). 1994a. *OECD jobs study, evidence and explanations, Part I: Labor market trends and underlying forces of change.* Paris: OECD.

OECD. 1994b. *OECD Jobs Study, Evidence and Explanations, Part II: The Adjustment Potential of the Labor Market.* Paris: OECD.

OECD. 1997. Economic performance and the structure of collective bargaining. *OECD Employment Outlook.* Paris: OECD.

OECD. 2004. *OECD employment outlook*. Paris: OECD.

OECD. 2006a. *OECD employment outlook*. Paris: OECD.

OECD. 2006b. *Society at a glance: OECD social indicators*. Paris: OECD.

Olson, M. 1990. *How bright are the northern lights? Some questions about Sweden (Crafoord Lectures)*. Lund: Institute of Economic Research, Lund University.

Salter, W.E.G. 1960. *Productivity and technical change*. Cambridge: Cambridge University Press.

Teulings, C., and J. Hartog. 2002. *Corporatism or competition? Labour contracts, institutions and wage structures in international comparison*. Cambridge: Cambridge University Press.

UN (United Nations). 2005. *Human development report*. New York: Oxford University Press.

Visser, J.J. 2006. Union membership statistics in 24 countries. *Monthly Labor Review* 129: 38–49.

World Bank. 1990. *World development report*. New York: Oxford University Press.

World Bank. 1995. *World development report*. Washington, DC: World Bank.

Labour Market Search

Dale Mortensen

Abstract

Time and other resources are required in the process by which workers and jobs are matched: this process is referred to as labour market search. Models of the search process have made contributions to our understanding of unemployment incidence and duration, labour turnover, earnings growth and wage dispersion. These models, which are based on the assumption that agents act in their own best interest, are designed characterize market equilibria in environments complicated by imperfect information and uncertainty. Consequently, they are also useful in the analysis of labour market policy.

Keywords

Beveridge curve; Labour market search; Law of one price; Market tightness; Matching; Productivity shocks; Reservation wage; Search costs; Search theory; Social networks; Unemployment; Unemployment insurance; Wage dispersion

JEL Classifications

J6

Labour market search refers to the process by which workers and employers find and match with one another in the labour market. The theoretical models that have been developed to understand the process explicitly account for the fact that search and matching are time consuming activities. The models have been used to interpret empirical data on phenomena that include unemployment duration and incidence, unemployment fluctuations, labour turnover, earnings growth, and wage dispersion and discrimination. As with other equilibrium models in economics, these are based on the assumption that participants in the labour market act in their own self-interest and that the market phenomena observed are explained by outcomes of the market participant interaction. Hence, they can and are used to address the consequence of labour market policy on labour market statistics and on the welfare of labour market participants.

Equilibrium search theory extends the standard competitive model of the labour market. In spite of the obvious usefulness and elegance of the competitive market equilibrium for the analysis of many questions, the framework in its simplest form excludes much of the phenomena of interest to a labour economist. To give but two examples, there is no unemployment in competitive equilibrium and workers with identical skills earn the same wage. Equilibrium labour market search theory was developed to explain these facts as well as other phenomena having to do with the dynamics of the employment and earning experiences of individual workers that cannot be accommodated in the standard model.

More generally, the labour market search framework has proven to be a useful tool for thinking about markets with 'friction', those that function without the market clearing auctioneer invoked in the competitive market framework. How do the participants in such markets come together? How are prices and the quantities exchanged at these prices determined? Search theory attempts to answer these questions.

Modelling the fact that the labour market experiences of individual workers take place in real time is an essential ingredient of the labour market search approach. After workers leave school and enter the labour market, most spend time seeking a job. Once employed, young workers seem to 'job-shop' by trying several employers and occupations before settling down to an extended period of employment with one. Later, employment spells are punctuated by interruptions attributable to changes in the individual's desire for employment, on the one hand, and the termination of the worker's current job, on the other. It is fair to say that virtually all the recent theoretical treatments of these phenomena are based on labour market search theory, and that theory informs the interpretation of most empirical studies that focus on them.

Individual Search Behaviour

The Reservation Wage

Labour market search theory began as a model of how a worker might gather information about employment opportunities. In the real world, there are many sources of such information. Economists and sociologists distinguish between formal and informal search channels. Formal channels are information sources provided by market institutions such as newspaper advertisements, public employment services, private employment agents and the Internet. Informal channels include friends, relatives, and neighbours, anyone in the workers' extended social network. It is well known that most workers find their jobs through these informal channels, a fact that underlines the decentralized nature of the labour market.

The first important paper in search theory, by George Stigler (1961), was an attempt to formalize the economic problem posed by the need to gather information about trading opportunities in a non-auction market where different prices for the same or close substitutes can coexist. He modelled the problem as one of choosing the size of a sample of prices drawn randomly from the available set. Given that the agent would purchase the good from the lowest-priced seller in the sample and must pay a fixed cost per price sample, how many quotes should the buyer seek?

Although this well-known problem in sampling theory provides some interesting insights, it does not serve well as a model of job search. In that context, it is the worker's time rather than money that is the principal cost incurred. Furthermore, the length of time spent by an unemployed worker is an observable quantity that is measured in both survey and administrative data. The models focused on the duration of search, which were simultaneously introduced by McCall (1970), Mortensen (1970), and Gronau (1971), became the basis for further work on the subject.

The idea underlying these models is that the duration of job search by an individual worker is usefully viewed as a random variable with the length determined by the worker's decision to accept or refuse offers as they arrive. In other words, instead of gathering a sample of job opportunities and selecting the one most preferred as in Stigler's formulation, these authors argued that it was more realistic to think of the search process as sequential in time. Offers arrive one at a time and the unemployment period ends when the worker accepts one of them. As McCall (1970) pointed out, this is formally an optimal stopping problem in the theory of decisions under uncertainty. It is well known that a reservation strategy is optimal: accept the first offer above some critical value.

Formally, let $F(w)$ characterize the distribution of offers and suppose that the number of offers received in a time period of unit length is a Poisson random variable characterized by the arrival rate λ. Assume that the worker is a risk neutral with an indefinite future life span. When the distribution of wages is known, the optimal strategy is to accept the first wage offered above a *reservation wage*. In other words, the reservation wage, denoted as R, is the lower bound on the set of wages that are acceptable to the worker. Accepting employment at the reservation wage must just compensate for any income forgone by becoming employed, which one can think of an unemployment benefit denoted by b, plus the option of continued search. Formally, if the worker does not plan to search while employed, the reservation wage is the implicit solution to the indifference condition.

$$R = b + \lambda \int_{R}^{\overline{w}} \left(\frac{w - R}{r + \delta} \right) dF(w), \qquad (1)$$

where r is the rate at which worker's discount future income, δ is the rate at which the worker can expect to lose a job, and $\overline{w}$ is the upper support of the wage distribution (see Mortensen and Pissarides 1999a, 1999b for a derivation of the equation). The second term, the option value of continued search, is the product of the offer arrival rate and the expected present value of the future gains in income attributable to the possibility of receiving an offer in the future.

Empirical Application

Labour economists later exploited the empirical implications of the original stopping model. In

one of the first such papers, Ehrenberg and Oaxaca (1976) pointed out that the expected duration of an unemployment spell and the expected post-spell wage were both increasing in unemployment insurance benefit. Specifically, since a spell ends only when an offer is received and it is found acceptable, the hazard rate of the unemployment duration distribution is the product $\lambda[1 - F(R)]$. Hence, the duration distribution is exponential with expectation equal to the inverse of the hazard that is increasing in R. As the post-spell distribution of wages is the distribution of offers truncated on the left by the reservation wage, the expected post-spell wage,

$$\frac{E\{w|\, w \geq R\} = \int_R^{\overline{w}} w dF(w)}{1 - F(R),}$$

is increasing in R. Since the reservation wage, the solution to (1), is increasing in unemployment income b, the expected duration of an unemployment spell and the average wage earned once employed should both increase with the generosity of the unemployment benefit.

Subsequently, Kiefer and Neumann (1979) used the fact that the model specifies the form of the statistical likelihood function for the observed length of unemployment spells and accepted wages. Formally, for a sample of n completed spells of unemployment followed by employment at an observable wage, a set of pairs denoted by (t_i, w_i), $i = 1, ..., n$, the likelihood of the observed sample is given by

$$L = \Pi_{i=1}^{n} \lambda[1 - F(R_i)]e^{-\lambda[1-F(R_i)]} \times \left(\frac{F'(w_i)}{1 - F(R_i)}\right),$$

for a set of workers who all sample from the same wage offer distribution. In this equation, R_i is the reservation wage of worker i which varies with the entitled unemployment insurance payment as determined by Eq. (4). Given observed values of b_i one can estimate both the offer arrival rate parameter and the distribution of acceptable wage offers using this structure, at least in principle. For a review of the early empirical literature that uses the duration analysis approach to estimation and search theory to interpret the results see Devine and Kiefer (1991). Wolpin (1995) provides an excellent treatment of the structural approaches to estimation of decision theoretic search models.

Equilibrium Wage Dispersion

The wage dispersion assumed in the stopping formulation of the job search problem is obviously inconsistent with the 'law of one price' that characterizes competitive equilibrium. Idiosyncratic match productivity is the simplest way to justify the assumption that a worker's employment opportunities can be described by a distribution of alternative wages. Although this justification may be sufficient for the purpose, Rothschild (1973) asks the following question: are there reasonable conditions under which wage dispersion, different wages paid to workers of identical skill, exists in equilibrium?

Consider the following simple one-shot game. All workers are identical and the common value of their marginal product is p in every firm. Assume that each worker receives a finite sample of job offers, say of size n, chosen at random from the set of all offers. Since the worker has no future in this formulation, his or her best strategy is to accept the highest offer in the set provided that it exceeds the opportunity cost of employment, denoted above by b. Of course, positive gain from trade requires that $p > b$. Given this strategy, what will profit maximizing employers offer?

Suppose that $n = 2$, that is, every worker receives offers from exactly two different firms. Because each worker will accept only the higher of the two offers, any employer paying a wage strictly less than all the others will hire no workers. Hence, a strictly positive fraction of employers must offer the lowest wage in the market: denote it by $\underline{w}$. It follows that the expected profit earned is $\underline{\pi} = (1 - \alpha)\frac{1}{2}(p - \underline{w})$ where α is

the fraction of workers that receive a strictly larger offer. In other words, the other wage drawn by the worker must also be the smallest in the market, an event that occurs with probability $1 - \alpha$. If so, the worker chooses one of the two offers at random, with probability 1/2.

But it will always pay an individual employer to break ties by offering slightly more. That is, because the profit obtained by doing so is $\pi = (1 - \alpha)(p - \underline{w} - \varepsilon) > \underline{\pi} = (1 - \alpha)\frac{1}{2}(p - \underline{w})$ for all sufficiently small $\varepsilon > 0$ if $p > \underline{w}$, it follows that the smallest offer is $\underline{w} = p$ in any equilibrium. Hence, all offers equal the competitive equilibrium wage, $w = p$. Obviously, this argument holds for any value of $n \geq 2$.

The fact that Bertrand competition obtains when every worker receives at least two offers would seem to rule out wage dispersion. However, this conclusion is false because the price-gathering process embodies an information externality as Rothschild (1973) points out. Suppose for the sake of argument that the first wage quote is costless but there is a small cost of finding a second. In this case, all workers sampling twice is not a non-cooperative equilibrium strategy in the game of wage search. Namely, if all workers see two prices, there would be no dispersion as we have just shown. But, if there is no dispersion, no worker has an incentive to pay the cost of obtaining a second price quote. It follows immediately that a single common wage equal to the opportunity cost of employment, $w = b$, is an equilibrium, a result due to Diamond (1971). However, Burdett and Judd (1983) demonstrate that another equilibrium generally exists in which a fraction of the workers seeks two offers while the complementary fraction obtains only one. The equilibrium in this case can be characterized by a unique continuous distribution of wage offers.

The details of the Burdett–Judd argument are beyond the scope of this article. However, the reason why an equilibrium of a wage posting game of the kind outlined above can be characterized by a distribution of offers is easily understood in the context of the sequential search model outlined above extended to allow for search on-the-job as in Burdett and Mortensen (1998).

If search costs are not too large, it is obvious that the employed as well as unemployed workers have an incentive to search when wage offers are dispersed. Hence, in the model there will be two kinds of workers: the unemployed, who only see one wage offer at a time, and employed workers, who will be able to choose between continuing employment at the same wage or moving to alternative employment when the opportunity arises. In short, at any point in time a strict subset of the workers have two offers while another fraction has one.

Given search on-the-job, employers who pay more attract a larger fraction of applicants and suffer less turnover. This trade-off between wage and turnover costs provides the reason for dispersion. Formally, let $V(w)$ represent the value of meeting a prospective employee to an employer who pays wage w. It is the product of two terms, the probability that an applicant will accept the wage offered and the present value of the future stream of profit that the employer can expect to earn if he or she accepts.

Under the assumption that all the employed workers accept any wage above a common reservation value, R, but an employed worker accepts if and only if the wage offer exceeds that currently earned, the acceptance probability is equal to

$$A(w) = u + (1 - u)G(w)$$

where u is the unemployment rate and $G(w)$ is the fraction of workers who currently earn less than the wage offered, w. Given that job separations occur at rate δ for exogenous reasons and a worker will quit when ever a higher-paying job is located, the expected present value of future profit is

$$J(w) = \frac{p - w}{r + \delta + \lambda[1 - F(w)]}$$

where the production of λ, the rate at which the worker generates outside offers, and $1 - F(w)$, the probability that an alternative offer exceeds the worker's current wage, is the rate at which an employed worker can be expected to quit. Hence, the expected value of meeting a worker contingent on the wage offered is

$$V(w) = A(w)J(w) = \frac{[u + (1-u)G(w)](p-w)}{r + \delta + \lambda[1 - F(w)]}.$$

Because a higher wage increases the acceptance rate and reduces the quit rate, a trade-off between wage and turnover costs is evident in this relationship. It is natural to assume that each individual employer will choose the wage to maximize the expected present value of future profit, the function $V(w)$, given the wage offers of all the other workers. However, because all the employers are identical by assumption and the offer distribution $F(w)$ is endogenously determined by all their wage-setting decisions, it follows that profits must be both maximal and equal to the support of any equilibrium distribution. Furthermore, because unemployed workers accept all offers and an employed worker accepts an offer only if it exceeds that currently earned, the acceptance probability is the unemployment rate and the quit rate is the offer arrival rate λ at the lowest equilibrium offer, which is the common reservation wage R. Hence, the equilibrium distribution is the unique solution for $F(w)$ to the following equal profit condition:

$$\begin{aligned} V(w) &= \frac{[u + (1-u)G(w)](p-w)}{r + \delta + \lambda[1 - F(w)]} \\ &= \frac{u(p-R)}{r + \delta + \lambda} \\ &= V(R) \forall w \in [R, \overline{w}] \text{ where } F(\overline{w}) \\ &= 1. \end{aligned}$$

As no worker will accept employment at a wage below R and any wage offer above $\overline{w}$ yields less profit, this condition is also sufficient for profit maximizing.

Variations and extensions of this model have been used to study the link between wages, labour turnover, the return to education, discrimination, and the duration of employment spells. The approach has also proven to be a valuable tool for the analysis of firm data on employment and workers flows. Eckstein and van den Berg (2007) provide a review of the literature that uses the model as the basis for parameter estimation. See Mortensen (2003) for a more complete development of the theory and a review of the empirical applications of the approach.

Equilibrium Unemployment

Search and matching model of unemployment, those based on the original two-sided search models of Diamond (1982), Mortensen (1982) and Pissarides (1985) have focused on the time required to find employment. In this family of models, match rent exists after worker and employer meet because finding an alternative is costly. When worker and employer meet, they are assumed to bargain over their joint output. According to Nash (1950), the outcome of the bargaining problem yields a wage equal to the flow value of unemployment, represented by the reservation wage R, plus some share of the rent attributable to the current job–worker match, when search for an alternative partner is assumed to be the outside option. Formally,

$$w = R + \beta(p - R), \beta \in (0, 1) \qquad (2)$$

where p represents match output and the value share parameter β reflects the worker's relative 'bargaining power'. When all matches are identical, the wage is the same for all job–worker matches. Furthermore, one can show all matches are acceptable if and only if match product p exceeds the opportunity cost of employment b.

As wages are the same in all jobs in this model, there are no quits, which implies that the expected present value of the future profit attributable to employing a worker is $J(w) = \frac{p-w}{r+\delta}$. Hence, an employer has an incentive to create a job whenever $J(w)$ exceeds the cost of doing so. For example, if the cost of advertising a job opening is c and the advertisement will attract applicants at frequency η per period, then it pays to post a vacancy whenever the expected cost of filling it, c/η, is less than the expected return to doing so as represented by $J(w)$.

At this point, it may have occurred to the reader that the rate at which workers are matched with jobs, denoted above as λ, and the rate at which vacant jobs are matched with worker, η, above

must be related. In the literature these are determined by a *matching function*, a market relationship between the flow of matches that form, and the number of workers and jobs seeking, a match. In this view, the matching function, denoted as $M(u, v)$, is a kind of 'production function' that relates the match inputs to match output. Given this function, it follows that $\lambda u \equiv M(u,v) \equiv \eta v$ since λu and ηv are both equal to the total match flow in the aggregate.

It is natural to suppose that the matching function, like an aggregate production function, is increasing, concave and homogenous of degree one. There is now a relatively extensive empirical literature, reviewed by Petrongolo and Pissarides (2001), which for the most part confirms these assumptions. When they hold, the vacancy-filling rate $\eta = M(u, v)/v = M(u/v, 1)$ is decreasing function of the ratio of vacancies to unemployment. Hence, if the expected return to filling a vacancy exceeds the cost of posting it, more vacancies will be created, driving down the expected return. Under the assumption of free entry, then, the market equilibrium number of vacancies posted at any point in time satisfies the free entry condition

$$\frac{c}{\eta} = \frac{c\theta}{M(1,\theta)} = J(w) = \frac{p-w}{r+\delta} \qquad (3)$$

where the vacancy–unemployment ratio, $\theta = v/u$, is referred to as *market tightness*.

In summary, an equilibrium solution to the model is a wage, reservation wage, and market tightness triple (w, R, θ) that joint satisfies the Eqs. (1), (2), and (3). Finally, because existing jobs are destroyed at rate δ and the flow of unemployed workers who find jobs is $\lambda u = M(1, \theta)u$, the steady state value of unemployment rate, that which equate the flows in and out of the unemployment state, is

$$u = \frac{\delta}{\delta + M(1,\theta)}. \qquad (4)$$

In other words, unemployment tends over time to a steady state value that increases with the rate of job destruction, δ, and decreases with market tightness, θ. Furthermore, market tightness depends on the incentive to create new jobs, the profit an employer can expect to earn in the future after the match forms. From Eq. (3), it follows that labour productivity, represented by the parameter p, is a major contributor to that incentive. Indeed, a positive shock to p first increases vacancies and, consequently, market tightness. Over time unemployment falls in response until its new steady state value is realized as characterized in Eq. (4). As a consequence, shocks to productivity trace out a downward sloping relationship between vacancies and unemployment, known in the empirical literature as the Beveridge curve. The effect of productivity shocks on unemployment is amplified by the fact that the rate of job destruction, δ in the model, falls with p. This channel of influence is incorporated into an extended version of the formal model by Mortensen and Pissarides (1994).

The theory has clear implications for labour market policy. For example, unemployment insurance is common in all developed economies and is either enacted or under consideration in many developing countries. As unemployment benefit is income contingent on being unemployed, it can be represented in the model by the parameter b. From Eqs (1) and (2) it follows that any increase in the benefit will raise wages, though its effect on the worker's bargaining threat point. In turn, the increase in wages will decrease future expected profit which will lead to a reduction in vacancies and market tightness according to the free entry condition (3). These facts together with Eq. (4) imply that a higher unemployment insurance benefit will raise the steady state level of unemployment. By clarifying this mechanism, the theory has played an important role in the debate over labour market policy reform in Europe.

For an extensive discussion of the matching model of unemployment and its implications see Mortensen and Pissarides (1999a, 1999b), Pissarides (2000), and the recent review article by Rogerson et al. (2005).

Summary

The development of the search-theoretic approach to the analysis of labour markets has

focused on two different issues, wage dispersion and unemployment, and the models used in each case are not fully consistent with one another. For example, in one branch wages are set by the employer while in the other wages are the outcome of a bilateral bargain between worker and employer This and other specification differences are subjects of current theoretical and empirical research designed to collect the features of each approach that best explain all the phenomena of interest. The ongoing research, designed among other purposes to integrate the two approaches, is reviewed by Rogerson et al. (2005).

See Also

- ▶ Matching
- ▶ Microfoundations
- ▶ Search Models of Unemployment
- ▶ Search Theory
- ▶ Search Theory (New Perspectives)
- ▶ Social Networks in Labour Markets

Bibliography

Burdett, K., and K. Judd. 1983. Equilibrium price distributions. *Econometrica* 51: 955–970.

Burdett, K., and D.T. Mortensen. 1998. Equilibrium wage differentials and employer size. *International Economic Review* 40: 889–914.

Devine, T.J., and N.M. Kiefer. 1991. *Empirical labor economics: A search approach*. Oxford: Oxford University Press.

Diamond, P. 1971. A model of price adjustment. *Journal of Economic Theory* 3: 156–168.

Diamond, P.A. 1982. Wage determination and efficiency in search equilibrium. *Review of Economic Studies* 49: 217–227.

Eckstein, Z., and G.J. van den Berg. 2007. Empirical labor search. *Journal of Econometrics* 136: 531–564.

Ehrenberg, R., and R. Oaxaca. 1976. Unemployment insurance, duration of unemployment, and subsequent wage gain. *American Economic Review* 66: 754–766.

Gronau, R. 1971. Information and frictional unemployment. *American Economic Review* 61: 290–301.

Kiefer, N.M., and G.R. Neumann. 1979. An empirical job search model with a test of the constant reservation wage hypothesis. *Journal of Political Economy* 87: 89–107.

McCall, J.J. 1970. Economics of information and job search. *Quarterly Journal of Economics* 84: 113–126.

Mortensen, D.T. 1970. Job search, the duration of unemployment and the Phillips curve. *American Economic Review* 60: 847–862.

Mortensen, D.T. 1982. The matching process as a noncooperative/bargaining game. In *The economics of information and uncertainty*, ed. J.J. McCall. Chicago: University of Chicago Press.

Mortensen, D.T. 2003. *Wage dispersion: Why are similar people paid differently?* Cambridge, MA: MIT Press.

Mortensen, D.T., and C.A. Pissarides. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61: 397–415.

Mortensen, D.T., and C.A. Pissarides. 1999a. Job reallocation, employment fluctuations, and unemployment differences. In *Handbook of macroeconomics*, ed. M. Woodford and J. Taylor. Amsterdam: North-Holland.

Mortensen, D.T., and C.A. Pissarides. 1999b. New developments in models of search in the labor market. In *Handbook in labor economics*, ed. O. Ashenfelter and D. Card. Amsterdam: North-Holland.

Nash, J. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Petrongolo, B., and C.A. Pissarides. 2001. Looking into the black box: A survey of the matching function. *Journal of Economic Literature* 38: 390–431.

Pissarides, C.A. 1985. Short-run equilibrium dynamics of unemployment, vacancies and real wages. *American Economic Review* 75: 676–690.

Pissarides, C.A. 2000. *Equilibrium unemployment theory*. 2nd ed. Cambridge, MA: MIT Press.

Rogerson, R., R. Shimer, and R. Wright. 2005. Search-theoretic models of the labor market: A survey. *Journal of Economic Literature* 43: 959–988.

Rothschild, M. 1973. Models of market organization with imperfect information: A survey. *Journal of Political Economy* 81: 1283–1308.

Stigler, G. 1961. The economics of information. *Journal of Political Economy* 69: 213–225.

Wolpin, K.I. 1995. *Empirical methods for the study of labor force dynamics*. Luxembourg: Harwood Academic.

Labour Markets

R. Tarling

The view taken of labour in the economic system is fundamental to economic theory. Classical writings accepted that the conceptualization of labour was a major issue in constructing theory and

developed ideas in an area of the economics discipline which is now 'political economy'. But labour in these early writings was regarded largely in terms of the individual, who by his very existence was part of a social, institutional and political system. How then could economics ever come to be seen to be remote from, or at least distinct from, sociology and political science?

The simplest procedure by which theorists can achieve the distinction is to distinguish between labour input as a sequence of services performed and the individual within whom the ability to perform these services is embodied. The sphere of economics is then confined to the allocation of these services, taking the process of extraction of these services from individuals as an issue at the boundary of the discipline. In political economy, that process is crucial and perhaps best expressed in the Marxist construct of 'labour power'. But neoclassical theory draws the line elsewhere: services are traded in an open market, just like any commodity, and individuals are presumed to offer to that market well-defined labour services. There is no extraction problem because the offers are voluntary, rationally chosen by each individual according to his utility of work.

This is then the most straightforward conceptualization of a labour market in which the services are traded as any commodity without reference to the source of these services. Markets will clear because marginal productivity theory allows the user to price the services and relate price and quantity while the suppliers of the services have a price governed by their utility preferences for work and can similarly trade-off price and quantity.

There is no difficulty in this theoretical framework in allowing for differentiation in the type of labour services. Some services may be quite distinct from others, so that we can talk about totally independent markets for the different groups of services. Alternatively, services may be more or less substitutable, for example combined with different amounts of physical capital, so that employers may select between a variety of combinations of labour services of different types. However, a unique solution is generally guaranteed by a 'best practice' technology which uniquely defines the demand for services of each type.

The more labour services are differentiated, the greater is the likelihood that there will be distinct markets for different kinds of labour services. This increases the possibility of resource bottlenecks and either markets which do not clear or which will only clear at very low or zero prices. Over time, the problem is resolved by human capital theory whereby investment takes place in those services which are in excess demand. A distinction is often made between general skills and firm-specific skills. However, in this context, the only relevance of the distinction is that it determines the size of the market in which the services are traded.

All of the problems for theory begin when we attempt to link the labour services with individuals who will provide these services. Each individual is capable of producing a range of services, not all of which may be required simultaneously. Few of these services will be instinctive, and will have required some period of learning, through formal or informal training or by experience. Human capital theory treats this process of training and skill acquisition as investment by the individual in a capability which can be taken to the market place and traded. There is a parallel with machines, where technology is embodied and a capability is taken to the market. But there are also some differences. The machine embodies technology which, in theory, is freely and readily available and inputs which will by assumption be available from the market. The individual has his native ability, which he will attain, and goes to the market to purchase inputs in the form of education, training and experience. However, his ability to realize his worth and to access the inputs depend on social organization and not on economic organization. Yet it can be asserted that the market will provide: that is, by assumption social organization is at least neutral and at best supportative of the market provision.

A rather more important factor is the cost of maintaining the capability to provide services. When machines are purchased as assets, the owner accepts a certain rate of physical depreciation and is responsible for maintenance and repair.

An individual is rather more akin to a machine that is leased: a contract has to be negotiated for the responsibility for maintenance and repair, and for the rate of physical depreciation. Whether machines or individuals, there is an economically optimal rate of exploitation whereby the machine or individual survives to provide services at a later date.

Slavery would be the equivalent of a sale of assets but labour is accepted as being a lease contract. A feudal system was not as extreme as slavery, but embodied a contract which was somewhat disadvantageous to labour. On the other hand, under slavery the owner is wholly responsible for maintenance and reproduction of the provision of services, not the slave. The interpretation of labour as a leased asset in a market system is the other extreme, where labour itself is responsible for its own maintenance and reproduction. Thus, just as any commodity, labour has a supply price based on the costs of its own maintenance and reproduction.

There is a fundamental difference when it comes to the issues of getting the services performed. The ability to provide services is embodied in an individual and that individual contracts to provide the services over a particular period of time. So far there is no difference with a machine. However, the actual extraction of those services is a feature for which a machine has been designed: so long as the technical environment is appropriate, the machine will function according to its design. But an individual is not so easily switched on. It may be assumed that the individual can, and will, provide services voluntarily and to his full ability. That presumes a great deal about social organization. Individuals have free will, although they may be coerced, and attitudes and performance are heavily influenced by workplace and community relationships.

This leads into the major issue, that of collective behaviour. What distinguishes labour from other factors of production is that individuals can form groups based on common interests, common aims, common circumstances and common environments. These groupings may be transient or permanent, informal or institutional, and formed in the workplace or the community. Furthermore the groupings may be formed around conflicts of interest as much as commonality of interests.

It should be recognized that collective behaviour, as it impinges on the economic system, is not restricted to labour as a factor of production. It is an aspect of behaviour of all agents in the economic system, whether owners of natural resources, owners of purchased assets, employers and those responsible for the operation of institutions. The social relationships of the economic system are not simply a matter of the collective bargaining between labour and individual employers in individual workplaces but a matter of class, owner and employers groupings affecting not only the well-known aspects of collective bargaining but also capital markets, product markets and industrial structure.

Investment in human capital, the costs of maintenance and reproduction, and the process of extraction already pose problems for the application of a theory concerned only with the allocation of labour services. There have been many theoretical developments to cope with these problems, linking consumption and work, rational expectations and implicit contact theory. They do not however, cope with social relations as such and still leave collective behaviour as an imperfection in the market operation. That is, the individual is at best no more than a natural resource available for wealth creation when the system requires. However, unlike other natural resources, individuals only ‘lease’ themselves to the economic system so that not only are they conditioned by their social environment initially, they subsequentially continually interact with it.

Thus, social relationships are crucial not simply because they impinge on the way labour markets work, but also because they may play a major role in determining what labour markets actually are. Defining a labour market is not simply a question of selecting a group of homogenous services, with a given set of demands and supply. The question of the relation between services and the individual, the factors influencing the price of trading, and the factors affecting job definition and the access to jobs are all likely to redefine markets.

Multiple labour markets, which do or do not interact, are recognized in a number of theoretical formulations. Non-competing groups, occupational strata and technical skills may all underpin different labour markets, the issue of whether or not these markets would interact being determined by the nature of production relationships. One of the more popular versions of labour markets is that of the dual labour market. As originally conceived, a primary market was determined by technology and industrial organization, in which employers had a vested interest in creating labour markets with promotion ladders, investment in firm or technology specific skills and limited points of entry. Such 'labour markets' would be created at firm, industry or occupation level and would generate a limited stock of jobs with attractive terms and conditions of employment. The remainder of the labour force competed for less attractive jobs and mobility between sectors was determined by the stock of attractive jobs and personal characteristics of members of the labour force. There is a strong underlying technological determinism in the dual labour market framework, but it does recognize institutional forms and market (particularly oligopolistic) structures.

This view of labour markets does begin to address one of the most difficult aspects of labour markets because it recognizes a form of hierarchy in the employment structure. One of the properties of a market is that demands and supplies are taken to the market place for trading to take place whether this takes place continuously with the 'invisible hand' fixing the prices or at the beginning of each production/consumption period. However, one of the features of employment is that many individuals already have jobs and that they establish certain property rights in those jobs merely because they fill them. It is only the vacancies and new jobs which are offered to the market. Furthermore, these jobs which are vacant are numerically largely to be found in existing firms and industries in which particular labour conditions already exist. So, in practice, there are very few jobs which are offered in a totally unconstrained way. Equally, there are rarely new groups of labour coming onto the market: most are additional supplies of workers with differentiating characteristics whose existence tend to condition the market environment for new entry.

The major difficulty for a market interpretation is that these considerations suggest a division in the labour market rarely encountered elsewhere. New jobs and vacancies created by individuals leaving jobs which need to be filled are part of the determination of access to employment in jobs where the wage may or may not be predetermined. Within a firm or industry, wages are reviewed, individually or collectively, at far more frequent intervals than the jobs or their incumbents. On the other hand, vacancies are being created and filled all the time in the economy so that, potentially, there is a far more frequent review of wages in jobs at the margin.

The greater the hierarchical structure of the jobs the greater the importance to be attached to the process of accessing employment. Thus, wage negotiations for jobs at the margin (new ones and those changing hands) are more likely to be affected by considerations of access, so that who fills the jobs is just as important as how much they are to be paid in the job. On the other hand, the wage determination is likely to be of a conventional kind, with employers offering wages measured as 'value for money' in some sense and applicants having a supply price based on some wage aspirations related to consumption and the costs of maintaining and reproducing their 'labour power'.

Wage determination for those in employment is likely to have wider scope, in that it will be more concerned with issues of extracting the labour services and with the wider aspects of the environment in which that extraction takes place. This suggests a divorce between the filling (and incumbency) of jobs and wage determination which greatly weakens the concept of a market for labour.

The two principal difficulties in the conceptualization of a labour market are the weakness of the direct link between the supply of labour and its price (seen either from the employers or the employees side) and the impact of social group formation on the processes within the labour market. At best, the employer's demand could be expressed through a marginal productivity theory and labour supply through a theory of social

reproduction. But, even then, the employer's demand is influenced by social group formation, reflected through the evolution of institutions and industrial structure, and labour supply is influenced by social group formation, reflected through the evolution of institutions and household/community structures. These influences are dynamic in nature and continually restructure the definition of labour markets.

This raises major problems for theories of inflation. In a macroeconomic context, it may well be that all that is important is the total volume of employment demanded and supplied, and the average price of services employed. However, the processes by which those aggregates and averages are determined create a distribution of incomes between different groups of labour and households or social units, which in turn keep the distribution of incomes in a state of flux and hence the aggregates also in a state of flux. Similarly, the deployment of labour and the extraction of services forms a major input into the determination by productivity, and hence unit labour costs and prices. Prices also feed back into average real incomes. Changing prices, like the tax system, falls differentially on groups of labour or households, altering the distribution of real incomes and further complicating the analysis of the process of inflation.

There have been periods in economic and social history when structures have remained sufficiently stable for apparent separation of employment, wage determination and inflation theory. But analysis of periods of radical change have found existing theories wanting. The search for a general framework for labour markets must continue, but economists must remove their blinkers if they are to make a worthwhile contribution.

See Also

- ▶ Collective Bargaining
- ▶ Hierarchy
- ▶ Implicit Contracts
- ▶ Incentive Contracts
- ▶ Layoffs
- ▶ Primary and Secondary Labour Markets
- ▶ Segmented Labour Markets
- ▶ Trade Unions

Labour Markets in the Arab World

Ragui Assaad

Abstract

This article reviews the various arguments that have been advanced to explain the defining features of Arab labour markets, which can be summarised as: high youth unemployment, especially among young women and educated youth; oversized public sectors and small and anaemic formal private sectors; rapidly growing but highly distorted and low-quality educational attainment; and low (and stagnant) female labour force participation. While acknowledging the validity of most of these arguments, I argue that these defining features are attributable in large part to the specific nature of the region's political economy, and, in particular, to the legacy of the so-called 'authoritarian bargain' social contracts that have characterised state–society relations in the post-colonial era.

Keywords

Arab labour markets; Authoritarian bargain; Female labour force participation; Human capital; Labour market segmentation; Social contracts; Unemployment

JEL Classifications

I25; J21; J24; J31; J45; O53; P52

The defining features of Arab labour markets have been well documented for some time. They include oversized public sectors; high youth unemployment – especially among young women and educated youth; small and anaemic

formal private sectors; rapidly growing but highly distorted and low-quality educational attainment; and finally low (and stagnant) female labour force participation (Assaad 2014). A number of arguments have been advanced to explain one or more of these features, including both supply-side and demand-side arguments, as well as arguments about the functioning of the labour market itself. Supply-side arguments include the demographic pressures resulting from the region's pronounced 'youth bulge' (Assaad and Roudi-Fahimi 2007), the educational mismatch hypothesis (Galal 2002; World Bank 2013a) and the region's conservative social and religious norms, which limit women's employment and, more generally, their engagement in the public sphere (World Bank 2013b). Demand-side arguments include the distorting effects of natural resource rents on the structure of the economy (Assaad 2006; Ross 2008) and on macreconomic stability (World Bank 2013a), misguided structural adjustment policies (El-Hamidi and Wahba 2005; Chaaban 2010; ILO/UNDP 2012) and the non-competitive and rent-seeking nature of the region's private sector (Malik and Awadallah 2013; World Bank 2013a). Arguments related to the functioning of the labour market have stressed the rigidities brought about by labour market regulations and institutions (Angel-Urdinola and Kuddo 2010).

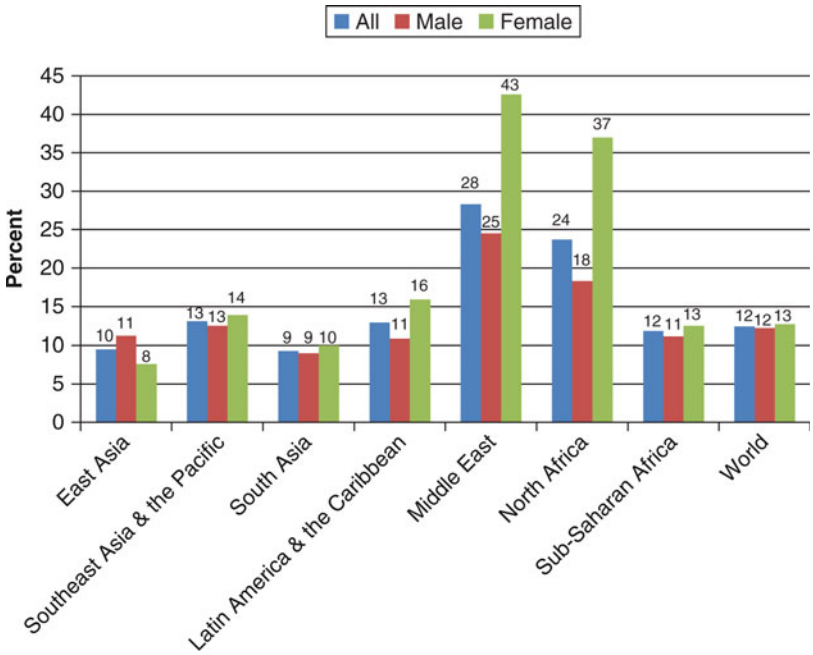
While most of these explanations have some degree of validity, I argue that the defining features of Arab labour markets are attributable in large part to the specific nature of the region's political economy, and, in particular, to the legacy of the so-called 'authoritarian bargain' social contracts that have characterised state–society relations in the post-colonial era (Desai et al. 2009; Amin et al. 2012). These social contracts, which are supported by the rentier nature of many Arab regimes, were based, in part, on the extensive use of public sector employment as a tool of political appeasement, thereby distorting the essential role of labour markets in the production and allocation of human capital. I argue that the demand-side distortions in the deployment of human capital brought about by politically driven public sector hiring have resulted, over time, in even more durable distortions in the supply of human capital, thus causing the observed skill mismatch in the labour market.

The first and most obvious consequence of using public sector employment to bribe politically sensitive groups into political quiescence is an oversized public sector and a strong preference for public sector work among new entrants. This results in high numbers of educated youth remaining unemployed while queuing for public sector employment, followed by the trapping of much of this human capital into unproductive public sector jobs. Facing barriers to employment in the private sector, educated women are typically even more concentrated than their male counterparts in the public sector, and thus more vulnerable to a curtailment of public sector hiring. They thus tend to remain unemployed longer and often withdraw from the labour force rather than take jobs in the informal economy that could jeopardise their marriageability.

The segmented labour market structure that results from the need to appease politically sensitive groups, in turn acts to shape in important ways the region's political economy. The fact that the vast majority of citizens in the Gulf monarchies, and the bulk of the middle class elsewhere, rely on the public sector for employment tends to foreclose the sort of class-based politics and class compromises that typically serve as the underpinning of capitalist democracies (Herb 2009). These influential groups will tend to have little stake in the vitality of the private sector and, in turn, the private sector sees no reason to give in to demands for higher wages and better working conditions. This dynamic is also reflected in the anaemic role of organised labour in most countries of the region. Trade unions, if they exist at all, almost exclusively represent public sector workers and therefore see their role as protecting the interests of these insiders rather than working to improve wages and working conditions in the economy as whole.

Defining Features of Arab Labour Markets

Although Arab economies vary greatly in terms of income level, hydrocarbon wealth and degree of



Labour Markets in the Arab World, Fig. 1 Youth unemployment rates by world region, 2012, ages 15–24 (ILO 2013)

labour abundance or scarcity, their labour markets share certain common features. Perhaps the best-known feature of Arab labour markets is the high rate of youth unemployment and the concentration of this unemployment among educated new entrants, especially female new entrants (Chaaban 2010; Assaad 2014). As shown in Fig. 1, youth unemployment rates in the Middle East and North Africa, as reported by the ILO, were by far the highest among all regions of the developing world, and this holds for both male and female youth. These two regions also have by far the highest ratio of female to male youth unemployment rates. The high youth rates are not simply because these two regions have high unemployment rates in general. The ratios of youth to adult unemployment rates are 3.8 and 3.4 in the Middle East and North Africa, respectively, which are significantly higher than the world average of 2.8, albeit lower than in South Asia and Southeast Asia and the Pacific, which have much lower youth and adult unemployment rates (ILO 2013, p. 108).

There is substantial evidence that youth unemployment in the Arab world is essentially a phenomenon involving educated new entrants searching for formal jobs, mostly in the public sector. Less educated workers tend to have much lower unemployment rates and tend to be employed primarily in informal jobs. Unemployment rates generally tend to increase with education, with a few exceptions like Palestine, where less educated males have been hard hit by the closure of the Israeli labour market to Palestinian labourers (Chaaban 2010; Assaad 2014).

The second most important feature of Arab labour markets is the disproportionate size of the region's public sectors, which continue to offer better compensation and working conditions than the private sector, leading to highly segmented labour markets. Nowhere is this segmentation as extreme as in the oil-rich Gulf countries, where most nationals are employed in the higher-paying public sector and where the private sector is almost entirely dependent on cheaper expatriate labour. The proportion of nationals employed in the public sector is as high as 92% in the UAE, 87% in Qatar, 86% in Kuwait and 72% in Saudi Arabia (Baldwin-Edwards 2011). Elsewhere, the public sector share of employment is also very high, reaching 54% in Iraq, 34% in Jordan and 27% in Egypt (World Bank 2013b).

The third dominant feature of Arab labour markets is the rapid increase in educational attainment, but the generally low quality of the education acquired. Of the 20 countries with the largest increase in average years of schooling from 1980 to 2010, eight were Arab countries (Campante and Chor 2012). At the same time, of the bottom 20 countries of the 148 countries ranked according to quality of basic education in the Global Competitiveness Report 2013/14, five were Arab countries (Schwab 2013). Fifty four per cent of students in the Middle East and North Africa (MENA) countries that participated in the Trends in International Science and Mathematical Study (TIMSS) 2007 scored below the 'low threshold' on the eighth grade mathematics tests, more than twice the international median of 25% scoring below that threshold (Mullis et al. 2008). (The worldwide average score is set at 500. The low threshold is set at 400.) The average score of eighth graders in both mathematics and science in all 15 Arab countries that participated in TIMSS 2007 was below the world average, and the average score in mathematics was below the low threshold in ten countries. The problem does not appear to be a question of limited resources, since some of the worst performing countries in both mathematics and sciences are among the oil-rich countries of the Gulf (Bouhlila 2011).

The final distinctive feature of Arab labour markets is the very low and relatively stagnant female labour force participation rate, despite rapid increases in educational attainment among women. Of the 20 countries with the lowest participation rates in 2011, 15 are Arab countries (World Bank 2014). Despite dramatic increases in women's educational attainment in most Arab countries, participation rates have remained low. The average female labour force participation rate in the Arab World increased by only three percentage points, from 20% in 1991 to 23% in 2011, compared to a world average of about 50% in 2011. These are the figures reported as the average female labour force participation rates for ages 15 + in the World Development Indicators database as modelled by the ILO (World Bank 2014). At this rate of increase, the World Bank estimates that it would take the region 150 years to attain the current world average (World Bank 2013a).

Common Explanations for These Defining Features

A number of explanations have been advanced to explain one or more of the defining features of Arab labour markets. Some of these explanations have stressed factors that affect the size and composition of labour supply, others have stressed structural features of Arab economies that limit and distort labour demand, and others have emphasised the workings of the labour market itself and the institutions that regulate it. The features that have received the most attention in the literature are the high rates of youth unemployment and the low female participation rates, although there is also considerable discussion of the reasons behind the poor quality of educational human capital in the region.

On the supply side, the severe demographic pressures resulting from the 'youth bulge' phenomenon are often cited as an explanation for the labour market insertion difficulties of Arab youth (Assaad and Roudi-Fahimi 2007; Chaaban 2010; El-Hamidi and Wahba 2004; World Bank 2004). While the effect of the youth bulge on swelling the ranks of new entrants to the labour market in the recent past is undeniable, it now clear that the youth bulge phenomenon will have peaked by 2015 in much of the region and that the share of youth in the population is either already falling or will begin to fall soon in the majority of Arab countries (Assaad and Roudi-Fahimi 2007). Exceptions include Iraq, West Bank and Gaza, Somalia and Yemen. The 'youth bulge' phenomenon is a feature that the region shares with other developing regions, such as South Asia and Sub-Saharan Africa, but it has had very different labour market manifestations in these regions, which generally do not suffer from high levels of youth unemployment (Assaad and Levison 2013). In fact, the experience of East Asia with the youth bulge showed that, as child dependency ratios decline as a result of a fall in fertility, the increasing proportion of youth of working age in the

population can potentially generate a demographic dividend if the additional human resources can be put to productive use (Bloom and Williamson 1998).

Another supply-side explanation for high levels of youth unemployment is the mismatch between the output of the education system and the needs of the labour market, leading to the low employability of graduates (World Bank 2013a). The mismatch is generally attributed to rigidities and inefficiencies within education systems that leave them unable to respond flexibly to signals from the labour market (Galal 2002; Muysken and Nour 2006; Salehi-Isfahani 2012; World Bank 2013a). Educations systems have traditionally been oriented toward the production of credentials suited for employment in the public sector rather than the skills demanded in an increasingly market-led economy, in what has been termed the 'credentialist equilibrium' (Salehi-Isfahani 2012).

The supply-side constraint on female labour supply resulting from conservative gender norms is perhaps the most common explanation brought up in the literature for the unusually low rates of female labour force participation in the Arab world (Sidani 2005; Spierings et al. 2010). Some scholars blame women's limited participation in the public sphere directly on 'Islamic culture' (Clark et al. 1991; Inglehart and Norris 2003), but such a position does not account for the wide variation in participation observed across the Islamic world. Others have attributed low female participation rates to social structures that emphasise women's modesty and reputational safety, and the primacy of the family and the domestic sphere in women's lives – what has been referred to as the 'gender system' (Miles 2002) or the 'traditional gender paradigm' (World Bank 2004; see also Youssef 1971, for an early elaboration of this perspective). Some authors blame the perpetuation of these patriarchal family structures and gender norms on the role of oil and oil-related revenues, which typically flow into male hands and allow them to perpetuate the traditional male breadwinner/female homemaker model (Karshenas and Moghadam 2001; Moghadam 2004a; Ross 2008). Yet others have explained the perpetuation of patriarchal family structures through the emergence of 'neopatriarchal' state institutions that resulted from the interaction between modernity and patriarchy in a context of dependent capitalism and state dominance, fuelled by oil-related revenues (Sharabi 1988; Moghadam 2004b; Haghighat 2005; Olmsted 2005; Charrad 2009). Given the very high rates of unemployment among young Arab women and the large wage penalties they incur in the private sector, a pure supply constraint argument is not tenable, at least for unmarried women. However, it could very well be that gender norms shape the kind of employment that is deemed socially acceptable for women, because of real or perceived risk of harassment or unrequited contact with men. This would lead to overcrowding of female labour into the few segments of employment that are deemed acceptable for women, such as health, education and the bureaucracy, leading to queuing for these jobs and a drop in wages in these feminised jobs (Assaad and El-Hamidi 2009; Assaad et al. 2014).

The role of oil and oil-related revenues, such as remittances, is also invoked to explain limited demand for female labour in Arab economies. One of the ways in which oil restricts demand for women's labour is through the 'Dutch Disease' phenomenon, whereby oil revenues appreciate a country's real exchange rate and thus alter the structure of the economy away from non-oil traded sectors, such as agriculture and manufacturing, and toward non-traded sectors, such as construction and services (Corden and Neary 1982). Since demand for women's labour tends to be more concentrated in these traded sectors, the Dutch Disease ends up reducing demand for female labour (Assaad 2006; Ross 2008). Ross (2008) makes the additional argument that oil resources reinforce neopatriarchal states. Ross's claims have generated a heated debate in the literature by proponents of the cultural values argument (Norris 2009; Groh and Rothschild 2012) and the social and kinship structures argument (Charrad 2009).

Demand-side arguments have also been invoked to explain the inadequacy of job creation in the formal private sectors of Arab economies

and the resultant high youth unemployment rates. These arguments range from the effects of structural distortions and macroeconomic instability brought about by high levels of dependence on mineral resources (World Bank 2013a), to the effects of misguided liberalisation and structural adjustment policies (El-Hamidi and Wahba 2005; Chaaban 2010; ILO/UNDP 2012), to slow-growing and uncompetitive private sectors characterised by cronyism, rent-seeking and insider privilege (World Bank 2013a; Amin et al. 2012; Malik and Awadallah 2013; Diwan et al. 2014). In effect these authors argue that Arab economies have simply lacked the necessary dynamism to create sufficient employment for the large and growing number of new entrants seeking employment. This lack of dynamism has been manifested in low rates of firm entry and exit as well as in low rates of growth for incumbent firms (World Bank 2013a). Because of regulatory barriers to growth in the firm space, private firms tend to be sub-optimally small and tend to stay that way. A number of reasons have been advanced for this lack of dynamism, but the most important of these has been the crony nature of capitalism in these countries, which determines the ability to make deals to avoid onerous regulations, privileged access to credit, and preferential access to government contracts and services (World Bank 2013a; Malik and Awadallah 2013; Diwan et al. 2014). In cases where employment creation has been adequate, as in Jordan and many of the Gulf countries, powerful private sector lobbies have ensured that access to cheap foreign labour was plentiful, so that most of the new jobs have gone to expatriate workers willing to work at much lower levels of compensation than nationals with similar skill levels.

A final set of explanations has focused on the functioning of the labour markets themselves and the nature of the institutions that govern them. A recent review of labour regulations in MENA carried out by the World Bank concluded that 'labour regulations, among other factors, introduces restrictions to employability in MENA' (Angel-Urdinola and Kuddo 2010). The assessment is based on employer responses in the World Bank Investment Climate Assessment (ICA) surveys, where firms in Egypt, Lebanon, Oman and Syria perceived labour regulations to be a major constraint on expanding hiring. The assessment also concludes that although hiring regulations are in line with international standards, firing regulations in non-GCC countries are quite strict compared to international benchmarks. The cost of realising terminations due to redundancy include the need to pursue complex administrative procedures, stipulations relating to reassigning and retraining workers, notice requirements, severance pay and other penalties (*ibid*.). A series of simulations based on Computable General Equilibrium models have concluded that reductions in payroll taxation on unskilled labour is a powerful instrument in promoting long-term unskilled employment (Agénor et al. 2007).

However, given the high levels of employment informality, even among formal firms, it is not clear that labour regulations are in fact a constraint on overall employment or simply on the degree of formalisation. Angel-Urdinola and Kuddo (2010) do in fact concede that enforcement of labour regulations remains weak in most countries and that compliance with firing regulations is limited outside the public sector. Achieving flexibility by not enforcing labour regulations is not optimal, as it leads to uncertainty, high degrees of informality and segmentation of the labour market between insiders (those with protected jobs) and outsiders (those on fixed-term contracts or hired informally) (*ibid*.). Nevertheless, as I argue below, the greatest distortion introduced by government in the operation of labour markets in Arab countries is most likely due to their role as employer rather than as regulator. By offering conditions of employment that significantly exceed what is available in the private sector and by being a significant employer, the government segments the labour market, encourages queuing for public sector jobs, and raises the 'reservation working conditions' of new labour market entrants. Because public sector wages, at least for men, are often comparable to those in the private sector, public wage setting itself is not the source of segmentation, but rather the other employment conditions associated with public jobs, such as job security, level of effort required and other employment benefits. These

generally far exceed those offered in much of the private sector, leading to raised expectations about minimum working conditions among job seekers, a notion that could be termed 'reservation working conditions' (Dougherty 2014).

The Role of Politically Driven Public Sector Hiring

While I do not discount the validity of most of the explanations outlined above, I argue that a number of the defining features of Arab labour markets are ultimately attributable to the specific nature of the region's political economy and, in particular, to the dominant, albeit eroded, social contract that has defined this political economy in the post-colonial era. Termed the 'authoritarian bargain' the basic tenets of the dominant social contract are that citizens are to accept political exclusion and disenfranchisement in exchange for employment in the public sector, free education, subsidised housing and health care, food subsidies and other benefits proffered by mostly rentier states (Desai et al. 2009). The most important of these benefits is an implicit promise of employment in the public sector for politically significant groups at relatively high wages, generous non-wage benefits and lifetime job security guarantees (Amin et al. 2012). These politically significant groups include secondary school and university graduates in countries such as Algeria, Egypt and Tunisia, specific sects, tribes and clans that are critical to the political survival of the regime in countries such as Iraq, Jordan, Syria and Yemen, and virtually the entire citizenry in the oil-rich Gulf monarchies, whose private economies rely almost entirely on cheap expatriate labour.

Oil revenues and other sources of rent, such as foreign aid flowing to the respective regimes, were critical in financing these authoritarian bargains and the politically driven public sector hiring that accompanies them (Ali and Elbadawi 2012). With the decline in oil prices and other sources of rent in the mid-1980s, fiscal pressures forced many of the region's regimes to move toward a more market-oriented model, but this did not necessarily signify a wholesale renegotiation of the authoritarian bargain. Many of the regimes simply resorted to an erosion of the regime's side of the bargain by limiting their provision of subsidised public services and curtailing their public sector hiring, while making sure to protect the entitlements of incumbents (Amin et al. 2012). This resulted in a system increasingly characterised by segmentation and an insider–outsider structure, whereby those who had already obtained an advantage, such as lifetime public sector employment, got to keep it, while eligible newcomers find it increasingly difficult to obtain the same. In time, the burden of adjustment was accumulating on the backs of an increasingly restless younger generation. With the recovery of oil prices in the 2000s, many oil-rich regimes, such as the Gulf monarchies and Algeria, revived the authoritarian bargain and resumed wholesale public sector hiring.

The liberalisation episodes were undoubtedly accompanied by a reduction of the role of the state in the economy in many instances, but rather than creating open, competitive market systems, partial reforms gave rise to the emergence of a crony capitalist class that could capitalise on their close relations with authoritarian rulers to engage in rent-seeking activities and concentrate economic gains into a few hands (Amin et al. 2012; Malik and Awadallah 2013). As a result, the formal private sectors in most Arab economies are weak and dependent, contributing little to employment growth. Sheltered from competitive pressures and faced with poor labour market information they can afford to indulge in preferential hiring practices that rely on social connections and personal networks rather than meritocratic principles (World Bank 2013a).

The combined effect of politically driven hiring practices in the public sector and non-meritocratic hiring in uncompetitive formal private sectors was the perpetuation of a highly dualistic labour market structure, characterised by a group of insiders with access to good jobs and a growing group of outsiders relegated to low-quality employment in the informal economy (for men) or to non-participation (for women). Using a simple Harris–Todaro model, it can be easily shown that when labour markets are

dualistic there is an incentive for those with a positive probability to obtain jobs in the favoured sector to queue for such jobs by remaining unemployed (Assaad 2014). This could potentially explain the high rates of unemployment among those with the threshold level of education to be eligible for government employment (typically secondary education or above), and the low level of unemployment for the lesser educated, who simply have no chance to obtain formal work and therefore do not search for it.

The significant advantages that come along with a public sector job and the closed nature of formal private employment has fed a continued strong preference for public sector employment on the part of job seekers, despite the declining probability of obtaining such employment in many contexts. This preference for public employment, in turn, drives human capital investments toward credentials that are deemed necessary to qualify for a public sector job, irrespective of the quality of education these credentials represent or the skills they impart (the so-called credentialist equilibrium). Thus, the demand-side distortion brought about by excessive public sector hiring at higher than market-clearing conditions of employment, when applied long enough, ends up introducing distortions in the supply side of human capital. This dynamic would explain the observed combination in the Arab world of rapid increases in educational attainment, poor educational quality and the skills mismatch.

While public sector hiring may not explain the low rates of female participation in the Arab world, it does explain the relative stagnation in these rates despite rapidly closing gender gaps in education and a traditionally strong gradient between educational attainment and labour force participation for women. This strong gradient between educational attainment and participation for women can be easily shown to result from the availability of government employment for educated women (Assaad 2014). With the slowdown in government hiring in recent years in countries such as Egypt, Jordan and Tunisia, educated women are increasingly faced with the option of either joining the informal economy or staying at home, with many opting to do the latter. Even unmarried women, who may not face a daunting domestic work burden, may find that jobs in the informal economy do not meet their reservation working conditions and may therefore prefer to remain unemployed. Such jobs may pose reputational risks that could potentially threaten a woman's marriageability. After marriage, women generally find informal wage work even less desirable, as it could potentially conflict with their significantly expanded domestic burdens (Hendy 2010; Assaad and Hendy 2013).

Conclusion

I have argued that Arab labour markets are characterised by certain well-established features, such as high unemployment, especially among educated new entrants, oversized public sector employment, rapid rates of educational attainment, but with low educational quality, and low and stagnant female labour force participation rates. I have reviewed a large number of explanations that have been brought forth in the literature to explain these common features. These include supply-side explanations, such as the youth bulge, rigid education systems and patriarchal gender norms and institutions; demand-side explanations, such as the distorting effects of resourcebased rents and non-competitive crony-based private sectors; and ones that attribute these features to regulatory and institutional features of Arab labour markets. While accepting many of these explanations as valid, I argued that many of the characteristic features of Arab labour markets are consequences of the use of public sector employment as a tool of political appeasement in the context of the dominant authoritarian bargain social contract. In making this argument, I do not discount the role of natural resource rents, but simply propose a different mechanism through which they might operate. The mechanisms highlighted in the literature include the role of natural resource rents in reducing labour supply, either by supporting private patriarchal norms or by allowing the perpetuation of neopatriarchal state institutions and their role in distorting labour demand through the Dutch Disease phenomenon.

Here I am emphasising, instead, their role in financing the fiscal costs of authoritarian bargains for what are essentially rentier states. I am also highlighting the uncompetitive and rent-seeking nature of the region's private sectors, which can also be linked to natural resource rents. These weak and dependent private sectors have been unable to exhibit the necessary dynamism to assume the mantle of job creation from public sectors increasingly unable to fulfil their part of the bargain due to rising fiscal pressures. These specific aspects of the region's political economy contribute directly to a highly segmented labour market structure that not only misallocates human capital in unproductive government jobs and produces unemployment queues, but also encourages investment in the wrong kind of human capital.

While I argued that the segmented nature of Arab labour markets is a product of specific aspects of the region's political economy, namely its reliance on authoritarian bargains funded by natural resource rents, it is also important to note the political economy consequences of this segmentation. As Michael Herb (2009) has argued, the almost exclusive reliance of Gulf citizens on the public sector for employment, and the similarly universal reliance of private sector employers on cheap foreign workers with no political rights, creates a peculiar kind of class conflict in which the basic compromises that underlie class politics in democratic societies can simply be avoided. Rather than fight for their share in capitalist profits, citizens in these countries have an incentive to mobilise to obtain a higher share of the oil rent in the form of higher government salaries, subsidised services and other forms of public welfare. Private employers, among whom ruling families are usually well represented, also have an incentive to engage in rent-seeking, but also to ensure that they retain access to a steady stream of cheap foreign workers with whom there is no need to strike any sort of social bargain. A similar argument can be extended to countries where public sector jobs have traditionally been used to appease the middle class rather than the entire citizenry. These middle class interests would not see their future welfare as tied to a dynamic private sector economy, but rather to a continued expansion of the state sector and to the jobs it engenders. With their membership almost exclusively made up of public sector workers, trade unions in these countries do not perceive their role as negotiating on behalf of the entire working class, but rather as defenders of the interests and entitlements of public sector insiders. Thus, the segmented nature of Arab labour markets tends to foreclose avenues for a classbased politics that is supported by a diversified and dynamic private economy.

See Also

- ▶ Algeria, Economy of
- ▶ Dutch Disease
- ▶ Egypt, Economy of
- ▶ Gender Roles and Division of Labour
- ▶ Jordan, Economy of
- ▶ Libya, Economics of
- ▶ Oman, Economy of
- ▶ 'Political Economy'
- ▶ Primary and Secondary Labour Markets
- ▶ Rent
- ▶ Rent Seeking
- ▶ Structural Unemployment
- ▶ Syria, Economy of
- ▶ Tunisia, Economy of
- ▶ Unemployment
- ▶ Women's Work and Wages
- ▶ Yemen, Economy of

Bibliography

Amin, M., R. Assaad, N. Al-Baharna, K. Dervis, R. Desai, N. Dhillon, A. Galal, H. Ghanem, C. Graham, D. Kaufmann, H. Kharas, J. Page, D. Salehi-Isfahani, K. Sierra, and T. Yousef. 2012. *After the spring: Economic transitions in the Arab world*. Oxford: Oxford University Press.

Agénor, P.-R., M. Nabli, T. Yousef, and H.T. Jensen. 2007. Labor market reforms, growth, and unemployment in labor-exporting countries in the Middle East and North Africa. *Journal of Policy Modeling* 29: 277–309.

Ali, O., and Elbadawi, I. 2012. The political economy of public sector employment in resource dependent countries. ERF Working Paper No. 673, Economic Research Forum, Cairo, Egypt.

Angel-Urdinola, D., and Kuddo, A. 2010. Key characteristics of employment regulation in the Middle East and

North Africa. Social Protection and Labor Discussion Paper No. 1006, World Bank, Washington, DC.

Assaad, R. 2006. Why did economic liberalization lead to feminization of the labor force in Morocco and de-feminization in Egypt? In *Gender impacts of trade liberalization in the MENA region*, 12–22. Tunis: Center of Arab Women Training and Research.

Assaad, R. 2014. Making sense of Arab labor markets: The enduring legacy of dualism. *IZA Journal of Labor and Development* 3(6): 1–25.

Assaad, R., and Levison, D. 2013. Employment for youth – a growing challenge for the global economy. Background Research Paper Submitted to High-Level Panel on the Post-2015 Development Agenda.

Assaad, R., and F. El-Hamidi. 2009. Women in the Egyptian labor market: An analysis of developments, 1988–2006. In *The Egyptian labor market revisited*, ed. R. Assaad, 219–257. Cairo: American University in Cairo Press.

Assaad, R., and Hendy, R. 2013. On the two-way relationship between marriage and work: Evidence from Egypt and Jordan. 19th Annual Conference of the Economic Research Forum, 3–5 March 2013, Kuwait.

Assaad, R., R. Hendy, and C. Yassine. 2014. Gender and the Jordanian labour market. In *The Jordanian labour market in the new millennium*, ed. R. Assaad, 105–143. Oxford: Oxford University Press.

Assaad, R., and Roudi-Fahimi, F. 2007. Youth in the Middle East and North Africa: Demographic opportunity or challenge? MENA Policy Brief, Population Reference Bureau (PRB).

Baldwin-Edwards, M. 2011. *Labour immigration and labour markets in the GCC countries: National patterns and trends*. London: LSE Kuwait Programme on Development.

Bloom, D.E., and J.G. Williamson. 1998. Demographic transitions and economic miracles in emerging Asia. *World Bank Economic Review* 12(3): 419–455.

Bouhlila, D.S. 2011. The quality of secondary education in the Middle East and North Africa: What can we learn from TIMSS' results? *Compare* 41(3): 327–352.

Campante, F.R., and D. Chor. 2012. Why was the Arab world poised for revolution? Schooling economic opportunities, and the Arab Spring. *Journal of Economic Perspectives* 26(2): 167–188.

Chaaban, J. 2010. Job creation in the Arab economies: Navigating through difficult waters. United Nations Development Programme, Regional Bureau for Arab States.

Charrad, M.M. 2009. Kinship, Islam, or oil: Culprits of gender inequality? *Politics & Gender* 5(4): 546–553.

Clark, R., T.W. Ramsey, and E.S. Adler. 1991. Culture, gender, and labor force participation: A cross-national study. *Gender and Society* 5(1): 47–66.

Corden, W.M., and J.P. Neary. 1982. Booming sector and de-industrialisation in a small open economy. *The Economic Journal* 92: 825–848.

Desai, R.M., A. Olfsgard, and T. Yousef. 2009. The logic of the authoritarian bargain. *Economics and Politics* 21(1): 93–125.

Diwan, I., P. Keefer, and M. Schiffbauer. 2014. *On top of the pyramids: Cronyism and private sector growth in Egypt (mimeo)*. Washington, DC: World Bank.

Dougherty, C. 2014. The labour market for youth in Egypt: Evidence from the 2012 school-to-work transition survey. Silatech Workshop on the Future of Arab Youth, London Middle East Institute, SOAS, University of London, 24–25 June 2014.

El-Hamidi, F., and Wahba, J. 2004. *Why does the MENA region have such high unemployment rates?* (mimeo).

El-Hamidi, F., and Wahba, J. 2005. The effects of structural adjustment on youth unemployment in Egypt. 12th Annual Conference of the Economic Research Forum, 19–21 December 2005, Cairo.

Galal, A. 2002. The paradox of education and unemployment in Egypt. Working Paper No. 67, Egyptian Center for Economic Studies, Cairo, Egypt.

Groh, M., and C. Rothschild. 2012. Oil, Islam, women, and geography: A comment on Ross (2008). *Quarterly Journal of Political Science* 7(1): 69–87.

Haghighat, E. 2005. A comparative analysis of neopatriarchy and female labor force participation in Islamic countries. *Electronic Journal of Sociology* 1: 1–1.

Hendy, R. 2010. Rethinking time allocation of Egyptian women: A matching analysis. ERF Working Paper No. 526, Economic Research Forum, Cairo, Egypt.

Herb, M. 2009. A nation of bureaucrats: Political participation and economic diversification in Kuwait and the United Arab Emirates. *International Journal of Middle East Studies* 41: 375–395.

ILO/UNDP. 2012. *Rethinking econoic growth: Towards productive and inclusive Arab societies*. Beirut, Lebanon: ILO Regional Office of the Arab States and UNDP Regional Bureau for Arab States.

ILO. 2013. *Global employment trends for youth: A generation at risk*. Geneva: ILO.

Inglehart, R., and P. Norris. 2003. The true clash of civilizations. *Foreign Policy* 135: 63–70.

Karshenas, M., and V.M. Moghadam. 2001. Female labor force participation and economic adjustment in the MENA region. *Research in Middle East Economics* 4: 51–74.

Malik, A., and B. Awadallah. 2013. The economics of the Arab Spring. *World Development* 45: 296–313.

Miles, R. 2002. Employment and unemployment in Jordan: The importance of the gender system. *World Development* 30(3): 413–427.

Moghadam, V.M. 2004a. Patriarchy in transition: Women and the changing family in the Middle East. *Journal of Comparative Family Studies* 35: 137–162.

Moghadam, V.M. 2004b. Women's economic participation in the Middle East: What difference has the neoliberal policy turn made? *Journal of Middle East Women's Studies* 1(1): 110–146.

Mullis, I., M. Martin, and P. Foy. 2008. *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynn School of Education, Boston College.

Muysken, J., and S. Nour. 2006. Deficiencies in education and poor prospects for economic growth in the Gulf countries: The case of the UAE. *Journal of Development Studies* 42(6): 957–980.
Norris, P. 2009. Petroleum patriarchy? A response to Ross. *Politics & Gender* 5(4): 553–560.
Olmsted, J.C. 2005. Is paid work the (only) answer? Neoliberalism, Arab women's well-being, and the social contract. *Journal of Middle East Women's Studies* 2(1): 112–139.
Ross, M.L. 2008. Oil, Islam, and women. *American Political Science Review* 102(1): 107–123.
Salehi-Isfahani, D. 2012. Education, jobs, and equity in the Middle East and North Africa. *Comparative Economic Studies* 54: 843–861.
Schwab, K. 2013. The Global Competitiveness Report: 2013–2014, p. 456. World Economic Forum, Geneva, Switzerland.
Sharabi, H. 1988. *Neopatriarchy.* Oxford: Oxford University Press.
Sidani, Y. 2005. Women, work, and Islam in Arab societies. *Women in Management Review* 20(7): 498–512.
Spierings, N., J. Smits, and M. Verloo. 2010. Micro-and macrolevel determinants of women's employment in six Arab countries. *Journal of Marriage and Family* 72(5): 1391–1407.
Youssef, N.H. 1971. Social structure and the female labor force: The case of women workers in Muslim Middle Eastern countries. *Demography* 8(4): 427–439.
World Bank. 2004. *Unlocking the employment potential in the Middle East and North Africa: Toward a new social contract*. Washington, DC: World Bank.
World Bank. 2013a. *Jobs for shared prosperity: Time for action in the Middle East and North Africa*. Washington, DC: World Bank.
World Bank. 2013b. *Opening doors: Gender equality and development in the Middle East and North Africa*. Washington, DC: World Bank.
World Bank. 2014. *World development indicators data set*. Washington, DC: World Bank.

Labour Mobility in the European Union

Jonathan Portes

Abstract

This article describes trends in labour mobility within the European Union since the Treaty of Rome and the resulting economic impacts, particularly since the accession of ten new Member States in 2004. It concludes that, half a century after 'free movement' was first incorporated into the founding treaties of the European Union, it is finally beginning to become an important factor in European economic integration.

Keywords

European Union; Labour mobility; Migration

JEL Classifications

F15; F16; F22; J61

Background

The European Union (I will refer throughout to 'the EU' to mean both the EU and its predecessor bodies such as the EEC) was founded on four basic principles: free movement of labour, capital, goods and services: these 'four freedoms' were set out in the original Treaty of Rome, which spoke of the 'abolition, as between Member States, of obstacles to the free movement of persons' (European Commission 1957). The point was to promote economic integration, in the widest sense, within the European area, which of course has since expanded considerably, and now covers 28 countries with a population of over half a billion people.

While the primary driver may have been a desire to promote European integration for its own sake, the founders of the EU also believed that there were large economic benefits. In fact, economic theory is ambiguous as to whether factor mobility (in this context, the free movement of labour and capital) is a complement to or a substitute for free trade (the free movement of goods and services). In a standard Heckscher–Ohlin model, they are pure substitutes. Either free trade or factor mobility will increase the efficiency of resource allocation and will maximise overall welfare; it is not necessary to have both.

Similarly, capital mobility may in some circumstances be a substitute for labour mobility. But in more recent, and arguably more realistic, trade models the picture is much less clear (see

Venables 1999, for a review). As long as there are frictions, or increasing returns to scale, for example, free trade and factor mobility (of labour, capital or both) will have different impacts (normally both will increase welfare, although this is not necessarily the case, as it depends on the nature of the frictions). The general consensus among economists is that labour mobility, like trade, is welfare-enhancing, although there may be significant distributional effects. Ozden (2015) provides a useful summary of the consensus view.

However, while the economic case may be strong in principle, other free trade areas (for example the North American Free Trade Area) or even customs unions do not typically involve free movement of people. So, from a purely economic perspective, free movement was not a necessary part of the European project; it would have been possible to have a customs union and an integrated economic space without it; the decision to make it one of the founding principles was a political as well as an economic choice. Labour mobility was seen as complementary not just to the economic aspects of European integration, but to its wider political objectives.

Trends in Labour Mobility Within the EU Before 2004

The period from the late 1950s to the early 1970s saw strong economic growth in most of the EU. Demand for labour was strong and unemployment low. However, intra-EU labour mobility remained quite low, compared to, for example, the USA, although there were significant flows from Italy to other EU countries, especially France. Labour demand was therefore largely met by immigration from outside the EU, especially Turkish 'guest workers' in Germany, North African migrants to France and – although the UK was not yet an EU Member State – Commonwealth migrants to Britain. Koikkalainen (2011) briefly reviews this period.

The economic crisis of the 1970s led to a sharp reduction in labour demand, and most EU countries (including the UK, which was now a member) attempted to reduce labour migration, although family ties, and the unexpected if unsurprising reluctance of so-called 'guest workers' to return to their countries of origin meant that significant migration continued from outside the EU. However, intra-EU mobility remained quite low throughout the 1980s and 1990s. The accession to the Union of Spain and Portugal in 1986 did not change this; although they had traditionally been countries of emigration, both were emerging from dictatorship and EU accession led swiftly to rapid economic growth and ample domestic demand for labour.

The 1980s and early 1990s did see a renewed push for greater market integration, launched under the umbrella of the 'Single Market'. However, the Commission's 1985 White Paper, which identified obstacles to the Single Market and set out proposals to address them, devoted only one relatively anodyne page to free movement: the focus was very much on product markets (European Commission 1985). A 1992 Recommendation did set out the case for some degree of harmonisation of social protection, in order to facilitate free movement, but this had relatively little practical effect.

So by 2000, although increasingly economically integrated in terms of trade, only slightly over 1% of EU citizens lived in a country other than their country of birth, and the previous decade had seen only a very modest upward trend (European Commission 2014a). This reflected not just unwillingness to move between countries, but a more general lack of mobility, even within countries. Inter-regional mobility in the EU was considerably lower than in the USA (Decressin and Fatás 1995). The result was that region-specific economic shocks in the EU were absorbed primarily through changes in employment rates (the participation rate in particular), while in the USA they were mitigated by labour mobility.

The potential downsides of this lack of mobility, despite the formal right to free movement, became more salient as the EU moved towards monetary union. The standard theory of optimal currency areas suggested that the costs of giving up the exchange rate as an adjustment mechanism (as a consequence of entering into an economic

union) would be reduced if other adjustment mechanisms, in particular labour mobility, were able to operate (Mundell 1961). There was therefore considerable concern that the lack of labour mobility posed a threat to the efficient operation of the incipient monetary union; this debate is summarised in European Commission (2014a).

Partly in response to these concerns, the EU undertook a number of initiatives designed to turn 'free movement of workers' from a formal right to one that appeared a realistic prospect to EU citizens. In particular, the Free Movement of Citizens Directive (European Commission 2004) simplified, consolidated and considerably extended the right to free movement for EU citizens, not just to take a job, but to look for one, and to be accompanied by family members (including non-EU citizens) as long as those exercising free movement were not an 'undue burden'. This also extended to non-discrimination against EU citizens, except in limited and temporary circumstances, in the operation of the benefit system. Its significance was not fully appreciated at the time.

2004 and 2007: Accession

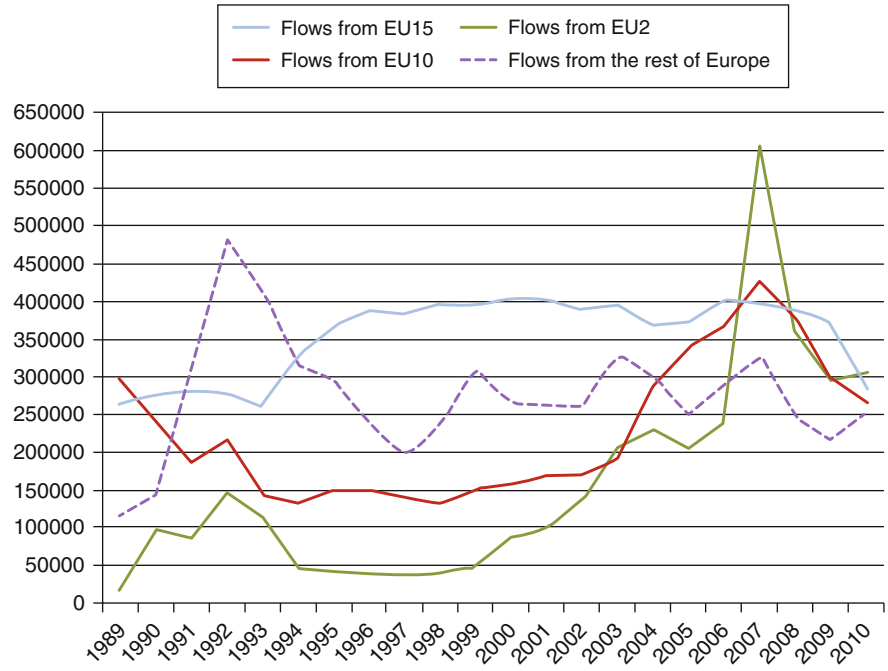
The accession, in May 2004, of ten new Member States, including a number of members of the former Soviet bloc (often referred to as the 'Accession 8', or A8, states – Poland, Hungary, the Czech Republic, Slovakia, Slovenia, Estonia, Latvia and Lithuania), radically changed the dynamic of intra-EU labour mobility. As set out above, free movement had originally (from an economic perspective; there were wider political motivations as well), been motivated by, first, theoretical arguments about optimal resource allocation; and, second, by its potential to serve as an adjustment mechanism in the face of asymmetric macroeconomic shocks. It had not been seen as operating in an area where there were very large, persistent, structural differences in wage levels, as was now the case.

Given these disparities, there was clearly a possibility of much larger intra-EU flows than had previously been the case. A number of Member States therefore took the opportunity under the accession treaties to impose 'transitional' restrictions on free movement of workers (the UK, Ireland and Sweden did not) for up to seven years. However, despite the restrictions, the impact of accession on intra-EU migration flows of both accessions was large and sustained, with substantial increases in migration to all the major economies of the existing EU, in particular the UK and Ireland. Indeed, Goodhart (2013) described the influx of A8 nationals to the UK as the 'biggest peacetime movement [of people] in European history'. While the largest flows were to countries without transitional restrictions, suggesting some diversion from those that did to those that did not, there were significant increases in flows to almost all existing EU member states. Given the questionable legal status (as regards employment) of migrants to countries that imposed transitional protections, official statistics may underestimate actual flows.

Holland et al. (2011) found that enlargement tripled A8 migration to the EU15, relative to a no-enlargement counterfactual. The main drivers were economic: Kahanec et al. (2014) found that migration responded both to structural economic differences between Member States and to short-term economic shocks, and that accession had led to significant increases in mobility, albeit hampered in part by the imposition of transitional restrictions. At an individual level, the vast majority of migrants moved to work, attracted by either higher wages or greater job opportunities. Location decisions were also influenced by cultural factors and network effects (Galgóczi et al. 2009).

In 2007, Bulgaria and Romania joined the EU; this too led to a significant increase in flows, although this time Spain and Italy were major destination countries. Transitional restrictions on Bulgarians and Romanians were finally lifted in all EU countries by 2014, so there is now complete free movement for the EU27 (some Member States still impose restrictions on Croatian nationals). Figure 1 shows the migration flows to EU15 countries.

Most recently, the Great Recession and the ensuing, and continuing, economic difficulties in some eurozone countries have also resulted in



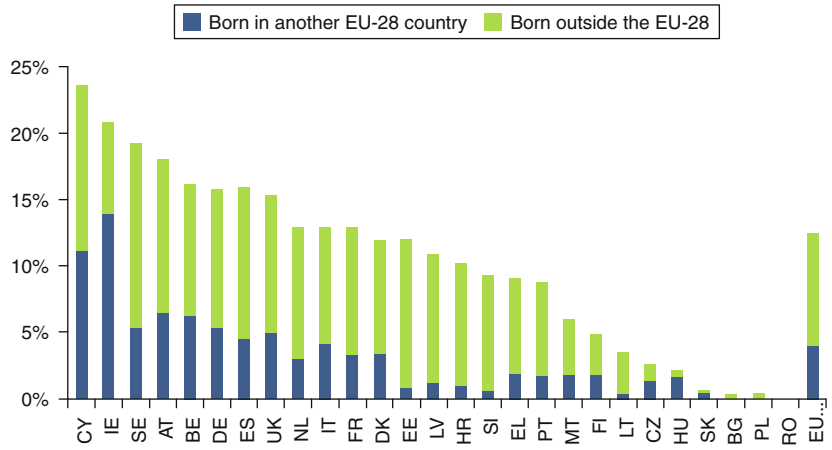
Labour Mobility in the European Union, Fig. 1 Migration flows to EU15 destination countries from Europe, by European regions of origin, 1989–2010 (source: Kahanec et al. 2014)

further changes in intra-EU migration flows. In particular, while Spain and Ireland have become less attractive to migrants both within and from outside the EU, out-migration has increased substantially from a number of countries where unemployment and/or youth unemployment is high: in particular Spain, Italy and Greece. Net migration from EU15 countries to the UK, historically quite stable, increased by about 50% between 2012 and 2014 (ONS 2014a). By contrast, emigration from Germany, France and the UK remains relatively low.

The result of these significant increases in intra-EU migration was that the proportion of EU nationals living in a Member State other than their birth country rose to about 3%. In a number of major countries, including Germany, the UK and Spain, it is now about 5%; in Ireland – traditionally a country of emigration rather than immigration – the figure is close to 15%. So while non-EU migrants still outnumber those from elsewhere in the EU in most Member States, there has clearly been a step change in intra-EU mobility. This rise has occurred at the same time as labour mobility within the USA seems to have fallen, perhaps as a result of changes to the labour and housing market (Beyer and Smets 2014); so while labour mobility is still a more important adjustment mechanism in the USA than the EU, the disparity is considerably less than 30 years ago. Figure 2 summarises the situation in 2013.

Macroeconomic and Labour Market Impacts

We now proceed to examine the economic impacts of this substantial increase in intra-EU migration (on both sending and receiving countries). As noted above, the primary motivation for migration was work, and most new migrants are in employment, with employment rates for intra-EU migrants well above rates for natives in most EU countries (European Commission 2013a). One notable feature of migrants from the new Member States was that, although they were not necessarily low-skilled, they primarily moved into



Labour Mobility in the European Union, Fig. 2 Share of working-age population born in other countries, 2013 (source: European Commission 2014a)

low-skilled employment in destination countries, and were concentrated in certain sectors (for example, construction, retail, hospitality, domestic work, food processing and agriculture) (Migration Advisory Committee 2014).

Standard theory predicts that a substantial movement of 'low-skilled' workers from relatively low-wage/low-productivity economies to higher wage/productivity economies will (assuming that the workers are employed in relatively low-skilled jobs) result in:

- increased output overall resulting from improved resource allocation
- increases in output in destination countries and reductions in sending countries – but impacts on per capita output will be considerably smaller, and possibly ambiguous
- reductions in the skill premium (the wage of a skilled worker relative to an unskilled one) and hence in wage inequality in sending countries and increases in the receiving countries, resulting from changes in the relative supply of skilled and unskilled labour; and (depending on labour market institutions) possibly reductions in unemployment in sending countries and increases in destination countries.

Simulation analysis with a large-scale econometric model largely confirms the expected macroeconomic impacts – broadly positive, but relatively small, and with some distributional impacts between countries (Holland et al. 2011) (see Table 1).

For labour markets, public and policy concern has focused on the distributional impacts – in particular potential negative impacts on employment and wages for low-skilled workers. Although the broad consensus in the economic literature is that the negative impacts of migration for native workers are, if they exist at all, relatively small and short-lived (see, for example, Constant 2014), much of this literature is US-based; given the perceived relative inflexibility of some European labour markets, policymakers were worried that negative impacts might be larger; and this concern was certainly shared by the public (Constant 2012).

There is now a considerable empirical literature on this topic, and the conclusions are surprisingly positive. Kahanec (2012) reviewing the literature, summarises: 'The pre-enlargement fears of free labour mobility proved to be unjustified. No significant detrimental effects on the receiving countries' labour markets have been documented, nor has there been any discernible welfare shopping. Rather, there appear to have been positive effects on EU's productivity'.

The largest number of empirical econometric studies focusing specifically on the labour market impacts of intra-EU migration have been conducted for the UK, beginning with Gilpin

Labour Mobility in the European Union, Table 1 Long-run impact on output before and after age adjustment – EU8 migration to EU15 countries (source: Holland et al. 2011). 'Age adjusted' simulations take account of the estimated age profile of migrants, which tends to increase the positive impact on receiving countries

	Long-run impact on GDP		Long-run impact on GDP per capita	
	Unadjusted	Age adjusted	Unadjusted	Age adjusted
Czech Rep	−0.20	−0.24	0.10	0.05
Estonia	−2.45	−2.98	−0.13	−0.65
Hungary	−0.33	−0.41	0.28	0.20
Lithuania	−4.89	−5.95	−0.29	−1.40
Latvia	−2.80	−3.32	−0.14	−0.69
Poland	−1.46	−1.75	1.00	0.70
Slovenia	−0.34	−0.40	0.00	−0.08
Slovakia	−1.92	−2.33	−0.09	−0.51
EU8	−1.25	−1.51	0.62	0.36
Belgium	0.28	0.36	−0.02	0.06
Denmark	0.42	0.56	−0.02	0.13
Finland	0.18	0.24	−0.10	−0.04
France	0.04	0.04	0.02	0.02
Germany	0.15	0.19	−0.02	0.02
Greece	0.07	0.08	0.03	0.04
Ireland	2.43	3.02	−0.79	0.01
Italy	0.12	0.15	−0.02	0.01
Neths	0.25	0.31	−0.03	0.04
Austria	0.30	0.39	−0.07	0.03
Portugal	0.06	0.06	0.04	0.04
Sweden	0.32	0.37	−0.08	−0.02
Spain	0.17	0.21	−0.04	0.01
UK	0.91	1.24	−0.13	0.20
EU15	0.33	0.43	−0.01	0.10

Source: NiGEM model simulation exercise

et al. (2006). A comprehensive literature review by the UK government (Devlin et al. 2014) found 'To date there has been little evidence in the literature of a statistically significant impact from EU migration on native employment outcomes'. For the other two countries that did not impose transitional provisions – Ireland and Sweden – there is, again, little evidence of significant labour market disruption (Doyle et al. 2006).

There are fewer studies directly measuring labour market impacts in other EU countries, but again those that there are suggest small or even positive impacts overall, resulting from complementarities. For example, Del Boca and Venturini (2014) find that many migrants to Italy work in the family care sector; given the inadequacy of state provision of family care, this helps to boost female labour force participation among natives. While this specific example is country- and sector-specific, there are likely to be other instances of this type of welfare-enhancing immigrant–native complementarity. Similarly, Farré et al. (2011) found a positive impact on high-skilled female labour supply in Spain.

There may also be other impacts on labour market institutions and structures, positive and negative, particularly if migration results in labour market segmentation (Migration Advisory Committee 2014). It is also possible that migration might impact on prices, particularly for goods and services produced in sectors in which migrant workers are concentrated. Frattini (2014) finds some, albeit inconclusive, evidence that immigration has reduced prices in low-wage service sectors. Moreover, as well as impacts on prices and wages, adjustment can also come via changes in

production technology (to reflect changes in labour supply): Dustmann and Glitz (2014) find some evidence of this in the German tradable goods sector.

So while, given data limitations, it is difficult to rule out some negative impacts (and it should be noted that there are relatively few studies using data during the Great Recession and the ensuing eurozone crisis), any such impacts seem to be quantitatively quite small, and outweighed by the broader positive impacts of improved resource allocation. While many European labour markets are far from healthy at present, especially for younger and low-skilled workers (European Commission 2014b), other economic and labour market developments (the general macroeconomic position, and the structural weaknesses of some EU labour markets) are likely to be far more important in explaining the problems.

What about the sending countries? Again, the evidence appears to be broadly positive. Zaiceva (2014) reviews country case studies and concludes (consistent with Holland et al. 2011) that out-migration has reduced unemployment and raised wages in sending countries, which also benefited from remittances. Micro-econometric evidence for Poland suggests that emigration raised wages overall, and in particular for intermediate-skilled workers (Dustmann et al. 2012). However, there are concerns about skill shortages. In Latvia and Lithuania, while emigration has clearly proved an invaluable safety valve during the crisis, helping to mitigate very high levels of unemployment, there is cause for concern about the longer-term demographic impact of emigration on an already ageing population with low fertility rates. Between 1990 and 2011, both countries saw their population fall by about 20% (*Lithuania Tribune* 2013).

'Benefit Tourism': The Welfare State Magnet Hypothesis

Despite the considerable evidence that migration flows were driven primarily by labour market factors (Kahanec et al. 2014), there is significant public concern in a number of EU15 countries that immigrants are attracted by the prospect of generous welfare benefits and are likely to become dependent on the state. The provisions of the Free Movement of Citizens Directive and ECJ case law, which oblige Member States to treat citizens of other EU countries comparably to their own citizens, may appear to facilitate such 'welfare' or 'benefit tourism'. This is particularly a concern in the UK (Duffy and Frere-Smith 2013), since the UK system gives some benefits (including in-work benefits for low-paid workers) on the basis of (income-related) 'need' rather than contribution (of course, contribution-based systems mean that new immigrants are unlikely to be eligible). However, such concerns are not confined to the UK; a recent, well-publicised German case at the European Court of Justice (2014) found that economically inactive EU citizens who go to another Member State solely in order to obtain social assistance may be excluded from certain social benefits.

It seems clear, however, that while there may, of course, be individuals who do indeed move between Member States to take advantage of the availability of social benefits, such a phenomenon is not quantitatively significant. A comprehensive compilation (European Commission 2013a) of the available data found that intra-EU migrants were substantially more likely to be in employment than natives, and significantly less likely than natives to claim disability or unemployment benefits. This was true for the UK as well as for other Member States, and the UK government was notably unable to substantiate its position that 'benefit tourism' was a significant policy concern (Portes 2013). The wider economic literature also supports the view that differences in benefit entitlements are not a significant driver of migration (Giuletti 2014).

Fiscal Impacts

Although intra-EU migration appears to be driven by differential labour market conditions rather than the relative attractiveness of welfare states, large movements of people can potentially have a significant impact on the public finances. In general, since intra-EU migrants, particularly those

from the new Member States, are significantly more likely to be in employment than non-migrants (European Commission 2013b), fiscal impacts might be expected to be positive. In addition, given their age profile, they are likely to place fewer demands on health services than natives (George et al. 2011). However, the magnitude of these positive impacts may be mitigated by the fact that many or most are in relatively low-paid employment.

Individual country studies have tended to confirm this: Dustmann and Frattini (2014) found that migrants from the EU to the UK made a significant positive contribution to the public finances, even during periods when the UK as a whole was running a fiscal deficit. Similarly, Ruist (2014) found that Bulgarian and Romanian migrants had made a substantial positive contribution to the Swedish public finances. Looking at four countries (Austria, Germany, the Netherlands and the UK), European Citizen Action Service (2014) found large positive fiscal contributions in all except the Netherlands (and then only if pensions were excluded).

Of course, it is hardly surprising that young migrants in employment make an initial positive fiscal contribution; proper assessment of fiscal impacts requires a life-cycle perspective (Preston 2014). In this context, there are various reasons to expect the impact to still be positive (in particular, migrants tend to arrive after they have left compulsory, publicly financed education). Lisenkova and Sanchez-Martinez (2013), using an overlapping generations model to project out the impacts of migration to 2060 for the UK, finds that migration has significant positive impacts on both GDP per capita and on the public finances over the very long term; a reduction in migration levels of 50% would require an increase in the tax rate on labour income of about 2% to preserve budget balance. This impact is particularly strong for intra-EU migrants, because of their young age structure and high activity rates. For Germany, Zentrum für Europäische Wirtschaftsforschung (2014), using a generational accounting approach, found that immigration overall made a significant contribution to restoring the long-run sustainability of the public finances – although the degree to which it did so depended crucially on the skill level of migrants.

However, positive net impact on public finances at the national level does not preclude significant impact on demand (and hence cost) at the local level, particularly if funding allocations do not adjust quickly (or at all) to reflect pressures resulting from migration (George et al. 2011). A notable recent example is the shortage of primary school places in some parts of the UK (especially London); this appears to be largely the result of poor planning on the part of central government, given the rise in the number of young children resulting from recent increases in migration (from both the EU and elsewhere).

Future Prospects

Many analysts thought that an initial surge of migrants from the new Member States was likely, given the very large wage differentials, as well as a natural inclination to take advantage of the new freedom to travel offered by the EU, but that net migration was likely to fall as economies converged and return migration increased, particularly after the Great Recession, which hit the EU15 more than the larger economies of the A8 (although the Baltic states suffered particularly badly). However, it is not clear that this is the case. Migration from the A8 to the UK and Germany remains significant. Moreover, while most A8 migrants are single at the point of arrival, many are now partnering and having children: more than 6% of all births in England and Wales in 2013 were to mothers born in the new Member States (ONS 2014b). Clearly, this makes them much more likely to settle for the long term, and greatly magnifies the longer-term economic and social impacts of intra-EU migration, compared to temporary labour migration. Migration flows are notoriously difficult to forecast (Mitchell et al. 2011); but it does seem likely that the 2000s saw a step change in the importance of intra-EU mobility to the EU economy.

Conclusion

The free movement of workers has long been a central principle of the European Union. But it was the accession to the EU of a number of former Eastern bloc states, with economies and labour markets very much weaker than those of many of the original Member States, which has made free movement an important economic phenomenon: for the first time, intra-EU labour mobility is becoming as important a driver, political and economic, of European integration as trade and capital flows.

While concerns about the impact of free movement (as well as immigration from outside the EU) in some countries have made it a major political issue, the economic impacts have been largely benign. There have been some, albeit relatively small, macroeconomic and fiscal benefits for destination countries, while most sending countries (with the possible exception of the Baltic states, over the longer term) have also benefited. Negative labour market impacts have been surprisingly small. 'Welfare tourism' appears to be a myth. And during and after the Great Recession, free movement appears to have operated as a useful channel of labour market adjustment. Overall, the natural optimism of economists over this experiment in labour market liberalisation appears to have been vindicated.

See Also

- ▶ Economic Demography
- ▶ European Labour Markets
- ▶ European Monetary Union
- ▶ European Union Single Market: Design and Development
- ▶ International Migration
- ▶ Labour Market Search
- ▶ Labour Supply
- ▶ Theory of Economic Integration: A Review
- ▶ Unemployment Measurement
- ▶ Unemployment

Bibliography

Beyer, R.C.M., and F. Smets. 2014. Labour market adjustments and migration in Europe and the United States: How different? *Paper presented at the 60th panel meeting of economic policy.* http://www.economic-policy.org/wp-content/uploads/2014/10/Beyer-Smets.pdf. Accessed 30 Mar 2015.

Constant, A.F. 2012. Sizing it up: Labor migration lessons of the EU enlargement to 27. In *European migration and asylum policies: Coherence or contradiction*, ed. C. Gortázar, C. Parra, B. Segaert, and C. Timmerman. Belgium: Bruylant.

Constant, A.F. 2014. Do migrants take the jobs of native workers? *IZA World of Labor.* doi:10.15185/izawol.10.

Decressin, J., and A. Fatás. 1995. Regional labour market dynamics in Europe. *European Economic Review* 39(9): 1627–1655.

Del Boca, D., and A. Venturini. 2014. *Migration in Italy is backing the old age welfare*. Discussion paper, 8328. Institute for the Study of Labor, Bonn.

Devlin, C., O. Bolt, D. Patel, D. Harding, and I. Hussain. 2014. *Impacts of migration on UK native employment: An analytical review of the evidence*. London: Department for Business Innovation & Skills . https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/287287/occ109.pdf. Accessed 30 Mar 2015.

Doyle, N., G. Hughes, and E. Wadensjö. 2006. *Freedom of movement for workers from Central and Eastern Europe. Experiences in Ireland and Sweden*. Stockholm: Swedish Institute for European Policy Studies . http://ec.europa.eu/enlargement/pdf/5th_enlargement/facts_figures/20065_en.pdf. Accessed 30 Mar 2015.

Duffy, B., and T. Frere-Smith. 2013. *Perceptions and reality. Public attitudes to immigration*. https://www.ipsos-mori.com/Assets/Docs/Publications/sri-perceptions-and-reality-immigration-report-2013.pdf. Accessed 30 Mar 2015.

Dustmann, C., and T. Frattini. 2014. The fiscal effects of immigration to the UK. *Economic Journal* 124(580): F593–F643.

Dustmann, C., and A. Glitz. 2014. *How do industries and firms respond to changes in local labor supply?* London: Centre for Research and Analysis of Migration . http://www.ucl.ac.uk/~uctpb21/Cpapers/CDP_18_11.pdf. Accessed 30 Mar 2015.

Dustmann, C., T. Frattini, and A. Rosso. 2012. *The effect of emigration from Poland on Polish wages*. London: Centre for Research and Analysis of Migration . http://www.cream-migration.org/publ_uploads/CDP_29_12.pdf. Accessed 30 Mar 2015.

European Citizen Action Service. 2014. *Fiscal impact of EU migrants in Austria, Germany, the Netherlands and the UK*. Brussels: European Citizen Action Service. http://www.epim.info/wp-content/uploads/2014/11/2BC_EU-migrants-final-2.pdf. Accessed 30 Mar 2015.

European Commission. 1957. *EU treaties*. http://europa.eu/eu-law/decision-making/treaties/index_en.htm. Accessed 30 Mar 2015.

European Commission. 1985. *Completing the internal market*. White Paper from the Commission to the European Council. Brussels: Commission of the European Communities. http://europa.eu/documents/comm/white_papers/pdf/com1985_0310_f_en.pdf. Accessed 30 Mar 2015.

European Commission. 2004. *Free movement of citizens directive 2004/38/EC*. Brussels: European Commission. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32004L0038&from=EN. Accessed 30 Mar 2015.

European Commission. 2013a. *Impact of mobile EU citizens on national social security systems*. http://ec.europa.eu/social/main.jsp?langId=en&catId=89&newsId=1980&furtherNews=yes. Accessed 30 Mar 2015.

European Commission. 2013b. *A fact finding analysis on the impact on the member states' social security systems of the entitlements of non-active intra-EU migrants to special non-contributory cash benefits and healthcare granted on the basis of residence*. http://ec.europa.eu/social/BlobServlet?docId=10972&langId=en. Accessed 30 Mar 2015.

European Commission. 2014a. *Labour mobility and labour market adjustment in the EU*. European Union Economic Paper 539. Brussels: European Commission. http://ec.europa.eu/economy_finance/publications/economic_paper/2014/pdf/ecp539_en.pdf. Accessed 30 Mar 2015.

European Commission. 2014b. *Employment and social developments in Europe 2013*. http://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=7684. Accessed 30 Mar 2015.

European Court of Justice. 2014. *Economically inactive EU citizens who go to another member state solely in order to obtain social assistance may be excluded from certain social benefits*. Press release, 11 November. http://curia.europa.eu/jcms/upload/docs/application/pdf/2014-11/cp140146en.pdf. Accessed 30 Mar 2015.

Farré, L., L. González, and F. Ortega. 2011. Immigration, family responsibilities and the labor supply of skilled native women. *B.E. Journal of Economic Analysis and Policy* 11(1): 1–46.

Frattini, T. 2014. *Impact of migration on UK consumer prices*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/328006/Impact_of_migration_on_UK_consumer_prices 2014.pdf. Accessed 30 Mar 2015.

Galgóczi, B., J. Leschke, and A. Watt. 2009. *EU labour migration since enlargement*. Farnham: Ashgate.

George, A., P. Meadows, H. Metcalf, and H. Rolfe. 2011. *Impact of migration on the consumption of education and children's services and the consumption of health services, social care and social services*. London: National Institute of Economic and Social Research . http://niesr.ac.uk/sites/default/files/publications/impact-of-migration.pdf. Accessed 30 Mar 2015.

Gilpin, N., M. Henty, S. Lemos, J. Portes, and C. Bullen. 2006. *The impact of free movement of workers from central and eastern Europe on the UK labour market*. Department of Work and Pensions, Working paper no. 29, Department of Work and Pensions.

Giuletti, C. 2014. The welfare magnet hypothesis and the welfare take-up of migrants. *IZA world of labor*. doi:10.15185/izawol.37.

Goodhart, D. 2013. *National citizen preference in an era of EU free movement*. Submission to the Government's 'Review of the Balance of Competencies'. http://www.demos.co.uk/files/DavidGoodhartSubmissionJuly2013.pdf. Accessed 30 Mar 2015.

Holland, D., T. Fic, A. Rincon-Aznar, L. Stokes, and P. Paluchowski. 2011. *Labour mobility within the EU – The impact of enlargement and the functioning of the transitional arrangements*. London: National Institute of Economic and Social Research.

Kahanec, M. 2012. *Labor mobility in an enlarged European Union*. Discussion paper 6485. Bonn: Institute for the Study of Labor. http://ftp.iza.org/dp6485.pdf. Accessed 30 Mar 2015.

Kahanec, M., M. Pytliková, and K.F. Zimmermann. 2014. *The free movement of workers in an enlarged European Union: Institutional underpinnings of economic adjustment*. Discussion paper 8456. Bonn: Institute for the Study of Labor. http://ftp.iza.org/dp8456.pdf. Accessed 30 Mar 2015.

Koikkalainen, S. 2011. *Free movement in Europe: Past and present*. Washington, DC: Migration Policy Institute. http://www.migrationpolicy.org/article/free-movement-europe-past-and-present. Accessed 30 Mar 2015.

Lisenkova, K., and M. Sanchez-Martinez. 2013. *The long-term economic impacts of reducing migration: The case of the UK migration policy*. Discussion paper no. 420. London: National Institute of Economic and Social Research. http://niesr.ac.uk/sites/default/files/publications/dp420.pdf. Accessed 30 Mar 2015.

Lithuania Tribune. 2013. *Opinion: Emigration and demographic decline in the Baltics*. 10 July. http://www.lithuaniatribune.com/44317/opinion-emigration-and-demographic-decline-in-the-baltics-201344317/. Accessed 30 Mar 2015.

Migration Advisory Committee. 2014. *Migrants in low-skilled work: The growth of EU and non-EU labour in low-skilled jobs and its impact on the UK*. London: Migration Advisory Committee . https://www.gov.uk/government/publications/migrants-in-low-skilled-work. Accessed 30 Mar 2015.

Mitchell, J., N. Pain, and R. Riley. 2011. The drivers of international migration to the UK: A panel-based Bayesian model averaging approach. *Economic Journal* 121(557): 1398–1444.

Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51(4): 657–665.

Office for National Statistics (ONS). 2014a. *Migration statistics quarterly report, November 2014*. http://www.ons.gov.uk/ons/dcp171778_386531.pdf. Accessed 30 Mar 2015.

Office for National Statistics (ONS). 2014b. *Country of birth of foreign born mothers*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/parentscountryofbirthenglandandwales/2014-08-28. Accessed 30 Mar 2015.
Ozden, Ç. 2015. A long commute. *Finance & Development* 52(1). http://www.imf.org/external/pubs/ft/fandd/2015/03/ozden.htm. Accessed 30 Mar 2015.
Portes, J. 2013. *Benefit tourism: The commission gives us some facts*. London: National Institute of Economic and Social Research.
Preston, I. 2014. *The effect of immigration on public finances*. London: Centre for European Policy Research . http://www.voxeu.org/article/immigration-and-public-finances. Accessed 30 Mar 2015.
Ruist, J. 2014. The fiscal consequences of unrestricted immigration from Romania and Bulgaria. *Working papers in economics*, no. 584. Department of Economics, University of Gothenburg.
Venables, A.J. 1999. Trade liberalization and factor mobility. In *Migration: The controversies and the evidence*, ed. R. Faini, J. De Melo, and K.F. Zimmerman, 23–48. Cambridge: Cambridge University Press.
Zaiceva, A. 2014. Post-enlargement emigration and new EU members' labor markets. *IZA World of Labor*. doi:10.15185/izawol.40.
Zentrum für Europäische Wirtschaftsforschung. 2014. *The fiscal effects of foreigners and immigration in Germany*. http://www.zew.de/en/news/2817/the-fiscal-effects-of-foreigners-and-immigration-in-germany. Accessed 30 Mar 2015.

Labour Power

G. de Vivo

The introduction of this notion has generally been regarded by Marxists as a crucial difference between their own and bourgeois economic theory. They have claimed that it allowed Marx to overcome a basic difficulty in Ricardo's (and, more generally, in classical) theory.

The most authoritative interpretation in the Marxist tradition, of the importance of the distinction between labour and labour-power, is the one given by Engels in his 1891 introduction to *Wage Labour and Capital*, where he argues that it could avoid the 'contradiction' into which 'economists' fell, when they 'applied the determination of value by labour to the commodity "labour"'. The contradiction would have been that for twelve hours' labour the worker receives as an equivalent value the product of six hours' labour. Either, therefore, labour has two values,. . . or twelve equals six! In both cases we get pure nonsense (Engels 1891, pp. 199–200).

But this nonsensical conclusion merely derives from a confusion between the value of labour (i.e. the wage) and the value of its product. No such confusion is to be found in Ricardo, or in those works of Marx of the 1840s (e.g. *The Poverty of Philosophy*, or the articles later republished as *Wage Labour and Capital*), where he had not made the distinction between labour and labour-power, and had simply treated labour as a commodity like anyone else.

Marxists have generally followed Engels's argument (see e.g. Mandel 1967, p. 81 ff.) and have accordingly failed to give a satisfactory explanation of the problem that the distinction was intended to solve.

The contradiction in Ricardo's theory, which he, according to Marx, had not even seen, can be formulated as follows. From the point of view of production of surplus value, materialized labour and living labour have different values. Indeed, surplus value . . . arises . . . from the fact that commodities or money (i.e., materialised labour) are exchanged for *more* living labour than is embodied . . . in them (Marx 1862–3, III, pp. 15–16). But Marx also notices that in Ricardo's theory the value of a commodity is equally determined by the quantity of *materialised* (*past*) labour and by the quantity of *living* (*immediate*) labour required for its production.

He therefore asks:

> If this difference [between *materialised* and *living* labour] is of no significance in the determination of the value of commodities, why does it assume such decisive importance when past labour (capital) is exchanged against living labour? Why should it, in this case, invalidate the law of value, since the difference *in itself*, as shown in the case of commodities, has no effect on the determination of value? Ricardo does not answer this question, he does not even raise it (Marx 1862–3, II, pp. 398–9).

Thus the problem is that 'labour has two values', as Engels had written, but in a sense

wholly different from the one envisaged by him: it has two values *with respect to materialized labour*. In the determination of the value of commodities it has the same value, in the capital/labour exchange it has a different value, than materialized labour.

The solution to this contradiction is provided by Marx in Chapter XIX, volume I, of *Capital*, where he soon faces the problem of explaining why 'the labourer . . . receives for 12 hours' labour . . . less than 12 hours' labour'. He notices that one cannot 'deduce the exchange of more labour against less, from the difference of form, the one being realised, the other living'. The solution he offers is the distinction between labour and labour-power:

> What the latter [the labourer] sells is his labour-power . . . Labour is the substance and immanent measure of value, *but has itself no value* (Marx 1867, pp. 502–3).

Thus, Marx seems to think that it is possible to escape the contradiction he had noticed, by distinguishing between 'labour' and 'labour-power', the former not being a commodity, but merely the 'substance' of value, which does not have itself any value. The problem of the relative value of living and materialized labour seemed therefore to disappear.

Marx had really seen a difficulty in Ricardo's theory which Ricardo had not seen – and one which had even been among the causes of the 'disintegration' of the 'Ricardian School'. The question whether 'accumulated labour' was more valuable than 'living labour', had in fact been the cause of many difficulties to the Ricardians, and had led to the abandonment of the labour theory of value. The difficulty however is not really overcome by Marx. The real issue behind it is in fact that of determining values by summing labours embodied at different times. The very fact that one must distinguish between 'antecedent' and 'present' labour in the capital/labour exchange – i.e. the very existence of profit – implies that one must also distinguish between 'antecedent' and 'present' labour when determining values. Marx's determination of the rate of profit, in his 'transformation of values into prices of production', is instead still based on the incorrect summing of labours of different dates (see also de Vivo 1982, p. 92 ff.).

See Also

- ▶ Abstract and Concrete Labour
- ▶ Labour Theory of Value

Bibliography

De Vivo, G. 1982. Notes on Marx's critique of Ricardo. *Contributions to Political Economy* 1: 87–99.

Engels, F. 1891. Introduction to K. Marx, wage labour and capital. In *Collected works*, vol. 9, ed. Marx Engels. London: Lawrence & Wishart, 1977.

Mandel, E. 1967. *The formation of the economic thought of Karl Marx*. London: New Left Books, 1971.

Marx, K. 1862–3. *Theories of surplus value*, vols. I–III. London: Lawrence & Wishart, 1969–72.

Marx, K. 1867. *Capital. A critique of political economy*, vol. I. London: Lawrence & Wishart, 1977.

Labour Process

William Lazonick

The labour process is a Marxian term that refers to the ways in which labour and capital combine to produce goods and services. The emphasis on the role of labour in the production process derives from Marx's (1867) distinction between labour-power and labour. Labour-power is the capacity to work that the capitalist purchases for a wage on the labour market; labour is the effort actually expended by a unit of labour-power in the production process. Given wages and prices, the surplus-value that the capitalist extracts from the production process depends upon the amount of labour services that he can elicit from the labour-power that he has purchased.

Based upon the distinction between labour-power and labour effort, Marx's theory of surplus-value analyses the generation of productivity and profitability within the capitalist

enterprise and concomitant impacts on the working conditions of the labouring population. Quite apart from the capitalist character of production, the transformation of inputs into outputs requires that human beings plan and execute the combination of their own productive capabilities with raw materials, tools, and machines. Within a complex social and technical division of labour, people invent processes, design products, build plant and equipment, coordinate various productive activities, handle tools, and tend machines.

Work occupies much of a person's active life, and can serve as a prime means of personal development. Marx argued, however, that capitalist control of the labour process tended to dehumanize the vast majority of workers. The social impact of capitalist development is, in his words,

> to mutilate the labourer into the fragment of a man, degrade him to the level of an appendage of a machine, destroy every remnant of charm in his work and turn it into hated toil, [and] . . . estrange from him the intellectual potentialities of the labour-process in the same proportion as science is incorporated in it as an independent power. (Marx, [1867], 1977, p. 799)

The only reward that the worker can hope to receive for his or her long hours of labour is a wage that just suffices for sustenance at a social acceptable standard.

Within the capitalist labour process, the alienating nature of work brings to the fore the conflict over the relation between effort and earnings. For a given wage, workers want to exert themselves as little as possible while capitalists want them to work as long and hard as possible. Marx's theory of surplus-value depends critically on the assumption that the capitalist has a degree of privileged access to the workers that he employs, permitting him to extract unremunerated effort from them.

Under competitive market assumptions, the capitalist takes all prices, including wages, as given, and technology is quickly diffused so that a particular capitalist cannot retain privileged access to process or product innovations for any appreciable period of time. But the worker's need to make a basic living, the deskilling of the labour force through technological change, and the existence of a homogeneous and hence interchangeable reserve army of labour all render the worker dependent on a particular capitalist employer. As a result, the capitalist is not entirely subject to the dictates of market forces in dealing with the worker in the labour process. The more dependent the worker is upon his or her particular employer, the more power the capitalist has to demand longer and harder work in return for a day's pay. The resultant unremunerated increase in the productivity of the worker per unit of time is the source of surplus-value.

Marx drew upon the historical experience of Britain's industrial revolution to develop his analysis of the labour process. He correctly emphasized the heavy reliance of the textile factories on the relatively low-waged labour of women and children, who, lacking social power to resist, were made to work long and hard (see Pollard 1965; Thompson 1967; Marglin 1974; Berg 1985; Lazonick 1986a). The hours of work were so extended that workers as well as more far-sighted members of the propertied classes organized for government legislation to limit the exploitation of labour. By the 1840s British factories were subject to a regulated working day, so that, for a given wage, exploitation within the labour process depended upon the amount of effort expended per unit of time rather than increases in the units of time that prolonged the working day.

Marx understood, also quite correctly, that the main obstacle confronting capitalists of the industrial revolution in attaining unremunerated intensification of labour was the resistance of skilled workers. In his view, the capitalist solution to worker opposition was the introduction of machinery into the labour process. According to Marx, machinery not only makes the capitalist less reliant on particular workers by superseding the strength and skills required of human beings in the labour process, but also displaces workers, adding to the reserve army of labour and rendering those who remain in employment all the more fearful of losing their jobs if they do not work long and hard enough. Marx recognized that, by overcoming the strength and skill limitations of humans, machinery is potentially *effort-saving*. But citing John Stuart Mill's contention that '[i]t

is questionable if all the mechanical inventions yet made have lightened the day's toil of any human being' (Marx [1867], 1977, p. 492), Marx argued that capitalists were able to use machinery as a powerful weapon against workers to increase effort levels and extract surplus-value.

In historical perspective, however, Marx misperceived the impact of technology in shaping the relations between capitalists and adult male workers in 19th-century Britain. Because of the limited managerial capabilities of relatively small firms in highly competitive industries, key groups of adult-male workers maintained considerable control over the technical division of labour, the flow of work on the shop floor, and the relation between effort and pay, even on the new technologies that gave rise to what Marx called 'modern industry'. During the long mid-Victorian boom, these workers consolidated their positions of job control as atomistic firms opted for collective accommodation with unions rather than let industrial conflict jeopardize the profits that were waiting to be made (Lazonick 1979; Harrison and Zeitlin 1985; Elbaum and Lazonick 1986).

So long as British industry dominated world markets, as it did in the last half of the 19th century, cooperation between capitalists and workers promoted productivity growth, permitting real wages to rise without cutting into profits. Organized workers who entered into stable relations with capitalists had the power to extract a share of productivity gains and could be enticed to invest in the development of specialized productive capabilities and work harder and longer for the sake of higher earnings.

In failing to see the sustained sources of power that key groups of British workers exercised over the relation between work and pay in the 19th century, Marx ignored the positive impact that cooperative industrial relations could have on productivity growth and the simultaneous increase in both wages and profits. He also overemphasized the deskilling of the labour force as a logical consequence of technological change, neglecting the ability of workers to influence the direction of technological change, both indirectly as new technologies were adapted to make use of available skills and directly as workers received training as technical specialists to develop and implement new technologies (Samuel 1977; Lazonick 1981, 1986b; Wood 1982; Lazonick and Mass 1984).

In Britain, shop-floor control persisted into the second half of the 20th century, and has only recently been challenged seriously by anti-labour policies and rapid deindustrialization of the Thatcher era. In historical perspective, Marx's analysis of the subjugation of labour to capital in the labour process would appear to be more applicable to the 20th-century experience of American capitalism, in which from the late 19th-century craft unionism and shop-floor control of the labour process were eradicated in the mass-production industries (Montgomery 1979; Brody 1980).

Indeed, Baran and Sweezy's (1966) influential analysis of US monopoly capitalism follows Marx in viewing the problem of surplus extraction as solved within the modern enterprise, focusing instead on the macroeconomic problems of surplus absorption. Integrating Marx's analysis of the British labour process of the 19th-century with the Baran and Sweezy analysis of US monopoly capitalism in the 20th-century, Braverman (1974) argues that degradation of work has remained the predominant social characteristic of the modern capitalist economy.

Braverman emphasizes the role of Taylorism or 'scientific management' – by which he means the separation of the conception of work within the managerial bureaucracy from the execution of work on the shop floor – in ensuring the triumph of capital over labour in 20th-century United States. He does not, however, recognize the vast development of skills among a considerable proportion of the labour force – albeit to a considerable extent on the part of workers who are segmented from shop-floor workers and integrated into managerial bureaucracies – required to operate within an evolving high-technology environment. Nor does he analyse how the divide between management and labour within the corporate enterprise – a phenomenon that occurred between 1880 and 1920 (Chandler 1977; Noble 1977) – enhanced capitalists' ability to extract unremunerated effort from workers.

In fact, as a method for increasing effort through piece-rate incentive schemes, Taylorism

was largely unsuccessful because 'scientific' managers sought to impose 'scientific' standards on non-unionized workers without giving them any assurance that they would share in the resultant productivity gains. In the absence of the countervailing power of craft unions that could bargain over the relation between effort and pay, management could be expected to subject workers to speed ups and stretch outs, while perhaps cutting piece rates, even in the presence of potentially effort-saving technological change. In response, even *unorganized* workers sought to restrict output by shop-floor solidarity to control the pace of work and defend themselves against unremunerated intensification of labour (Lazonick 1983 and 1984).

Braverman (1974, p. 85) argues: 'Logically, Taylorism belongs to the chain of development of management methods and the organization of labor, and not to the development of technology, in which its role was minor'. But to dissociate Taylorism from technology is to miss the essence of the problem that the managers of mass production faced. The movement towards what was called more generally 'systematic management' arose at a time when capitalists were making large fixed investments in new mass production technologies (Litterer 1963). The profitability of these investments depended upon the achievement of high rates of throughput, which would not be forthcoming if operatives saw fit to restrict output. Effort-saving technology held out the prospects for simultaneously lightening the physical strain of work and increasing productivity. But workers had to have some assurance that they would be able to appropriate a share of increased productivity if they were to cooperate in the actual generation of those gains (Lazonick 1984).

During the early decades of the 20th century, American capitalists searched for methods of labour management that would increase productivity without granting the workers any formal control over the determination of the relation between effort and pay. One widely used method was close supervision of the pace of work, but its success was limited during periods of prosperity by the ability of workers to exit from undesirable workplaces.

A complementary means of both reducing labour turnover and eliciting high levels of effort from workers was the offer of high wages – a method made famous by Henry Ford's five-dollar day, instituted in 1914 in conjunction with the introduction of the automated assembly line. There are those who see 'Fordism' as the ultimate achievement in mass production prior to the computer revolution begun in the 1970s (for example, Piore and Sabel 1984). In fact, Ford had only short-lived success with the high-wage strategy because, in the face of a growing used-car market and demand for more luxurious cars, the competitive advantage that the company had gained from mass producing the Model T slipped away. By the early 1920s, wages paid by Ford were no higher than his competitors, and the company had the worst labour relations in the industry (Chandler 1964; Meyer 1981; Hounshell 1984).

The longer-run solution to the problem of labour extraction was for corporations to hold out the promise of job security and upward mobility within the firm as the reward for hard and diligent work. During the 1920s, mass-production corporations instituted a dramatic change in labour relations as they began to make use of internal job ladders, not only within the burgeoning managerial bureaucracies but also among blue-collar workers. The erection of vertical job and wage structures represented a managerial strategy to discourage workers as individuals from seeking to better their lot by mobility via the external labour market. Instead workers who proved themselves dependable, loyal, and hardworking were offered opportunities for better work conditions, security, and pay within the firm (Slichter 1929; Lazonick 1983, 1986b; Jacoby 1984).

The effective use of internal job ladders is dependent on the growth of the firm. Internal job ladders will only induce hard work if employees observe that the higher level rungs of the ladders remain in place – a promise that many US mass production corporations could make by the 1920s by virtue of their oligopolistic market control. In turn, the ability of dominant firms to extend their market power was due in part to their ability to

deal with the problem of labour effort by strategies such as internalizing the labour exchange.

For dominant firms, a dynamic of rapid corporate growth was set in motion, only to be cut short as the Great Crash and its aftermath created macroconditions that even the corporate giants could not control. Unencumbered by debt, and hence immune from external pressures to produce at any cost, the response of the corporate mass-producers to the Great Depression was to cut back production and employment dramatically. The internal job structures erected in the 1920s collapsed. Significantly enough, IBM, a corporation that remains well-known for its permanent employment system, was able to keep its labour force fully employed during the 1930s by suplying 'business' machines to the expanding government sector under the New Deal (Sobel 1983, ch. 4).

During the 1930s, however, most large manufacturing corporations were unable to provide steady employment. Workers organized, the state intervened, and by the 1940s, workers had won seniority protection and the right to bargain over wage levels and differentials for management-determined job structures. Management had to share power over the determination of wage structures with unions. But the newly acquired union prerogatives meshed well with the strategy of erecting internal job ladders that the mass production corporations had been pursuing in the non-union era before the Great Depression. That strategy once again became viable in the 1940s, 1950s, and 1960s as the US economy entered a long boom characterized by expansion and diversification of the large corporations and American domination of world markets (Doeringer and Piore 1971; Edwards 1979).

In recent decades, however, the rise of international competition has made it more difficult for many US firms to promise job security and upward mobility to their workers. At the same time, the consolidation of social security systems has increased the level of the available 'social wage' and reduced the cost of job loss to many workers, making it more difficult for capitalists to enforce discipline in the workplace, with adverse impacts on productivity (Schor and Bowles 1984; Bowles 1985).

Recognizing the relation between alienated labour and low levels of effort, management has sought to deal with the problem of labour extraction by altering the technical and hierarchical division of labour in ways that 'humanize' work. These experiments often result in productivity increases on the shop floor. But, in the United States at least, they have been typically short-lived because, in redefining the hierarchical division of labour, the experiments inevitably infringe on traditional managerial prerogatives (Zimbalist 1975; Marglin 1979). In the late 20th century, American corporations are again searching for new methods of labour management that will yield profits without sacrificing hierarchical control.

A prime impetus for attempts to restructure the labour process in Western capitalist economies is the rise of Japanese competition over the past two decades. After World War II, many Japanese firms replaced militant labour unions by company unions that served to develop cooperative relations between labour and management characterized in part by vertical job structures that permit substantial mobility from the one into the other (see, for example, Cusumano 1985). Within dominant firms, internal job structures and permanent employment systems give many workers long-term stakes in the firm and assure them of shares of productivity gains. As a result of the integration of particular workers into the structure of the enterprise, Japanese managers can delegate authority over day-to-day decisions to workers into the shop floor without undermining hierarchical control much more readily than is the case in US or British firms, with apparently beneficial impacts on productivity.

The development of the labour process in dominant capitalist economies such as Britain, the United States, and Japan over the past century, therefore, reveals a quite different evolution of capital–labour relations from that envisioned by Marx. Exploitation of labour based upon highly intensified work for low pay certainly remains an important source of surplus-value in advanced capitalist economies. But, as research into

labour-market segmentation argues, such Marxian-type exploitation characterizes 'secondary', not 'primary' relations of production in modern capitalist economies (Gordon et al. 1982; Wilkinson 1981; Osterman 1984).

Marx's insights into conflicts of interest between capital and labour in the production process remain invaluable as points of departure for analysing the socioeconomic evolution of capitalism. The history of *successful* capitalist development demonstrates, however, the capacity for the economic system to transform conflict into cooperation so that, in fact, many if not most workers perceive that, in attacking institutions of private enterprise and accumulation, they may have much more to lose than their chains.

See Also

- ▶ Braverman, Harry (1920–1976)
- ▶ Capital as a Social Relation
- ▶ De-skilling
- ▶ Division of Labour
- ▶ Marxist Economics
- ▶ Taylorism

Bibliography

Baran, P., and P. Sweezy. 1966. *Monopoly capital*. New York: Monthly Review Press.

Berg, M. 1985. *The age of manufactures, 1700–1820*. London: Fontana.

Bowles, S. 1985. The production process in a competitive economy: Walrasian, Neo-Hobbesian, and Marxian models. *American Economic Review* 75(2): 16–36.

Braverman, H. 1974. *Labor and monopoly capital*. New York: Monthly Review Press.

Brody, D. 1980. *Workers in industrial America*. New York: Oxford University Press.

Chandler Jr., A.D. 1977. *The visible hand*. Cambridge, MA: Harvard University Press.

Chandler Jr., A.D. 1980. *Giant enterprise: Ford, General Motors, and the automobile industry*. New York: Harcourt, Brace & World.

Cusumano, M. 1985. *The Japanese automobile industry*. Cambridge, MA: Harvard University Press.

Doeringer, P., and M. Piore. 1971. *Internal labor markets and manpower analysis*. Lexington: D.C. Heath.

Edwards, R. 1979. *Contested terrain*. New York: Basic Books.

Elbaum, B., and W. Lazonick (eds.). 1986. *The decline of the British economy*. Oxford: Clarendon.

Gordon, D., R. Edwards, and M. Reich. 1982. *Segmented work, divided workers*. Cambridge: Cambridge University Press.

Harrison, R., and J. Zeitlin (eds.). 1985. *Divisions of labour*. Brighton: Harvester.

Hounshell, D. 1984. *From the American system to mass production, 1800–1932*. Baltimore: Johns Hopkins University Press.

Jacoby, S. 1984. The development of internal labor markets in American manufacturing firms. In *Internal labor markets*, ed. P. Osterman. Cambridge, MA: MIT Press.

Lazonick, W. 1979. Industrial relations and technical change: The case of the self-acting mule. *Cambridge Journal of Economics* 3(September): 231–262.

Lazonick, W. 1981. Production relations, labor productivity, and choice of technique: British and US cotton spinning. *Journal of Economic History* 41(3): 491–516.

Lazonick, W. 1983. Technological change and the control of work: The development of capital–labor relations in US mass production industries. In *Managerial strategies and industrial relations*, ed. H. Gospel and C. Littler. London: Heinemann.

Lazonick, W. 1984. Work, pay, and productivity: Theoretical implications of some historical research. Photocopy, Harvard University.

Lazonick, W. 1986a. Theory and history in Marxian economics. In *The future of economic history*, ed. A. Field. Hingham: Kluwer-Nijhoff.

Lazonick, W. 1986b. Strategy, structure, and management development in the United States and Britain. In *Development of managerial enterprise*, ed. K. Kobayashi and H. Morikawa. Tokyo: University of Tokyo Press.

Lazonick, W., and W. Mass. 1984. The performance of the British cotton industry, 1870–1913. *Research in Economic History* 9: 1–44.

Litterer, J. 1963. Systematic management: Design for organizational recoupling in American manufacturing firms. *Business History Review* 37(Winter): 369–391.

Marglin, S. 1974. What do bosses do?: The origins and functions of hierarchy in capitalist production. *Review of Radical Political Economics* 6(2): 60–112.

Marglin, S. 1979. Catching flies with honey: An inquiry into management initiatives to humanize work. *Economic Analysis and Workers' Management* 13: 473–487.

Marx, K. 1867. *Capital*, vol. I. New York: Vintage, 1977.

Meyer, S. 1981. *The five dollar day*. New York: State University of New York Press.

Montgomery, D. 1979. *Workers' control in America*. Cambridge: Cambridge University Press.

Noble, D. 1977. *America by design*. New York: Oxford University Press.

Osterman, P. (ed.). 1984. *Internal labor markets*. Cambridge, MA: MIT Press.

Piore, M., and C. Sabel. 1984. *The second industrial divide*. New York: Basic Books.

Pollard, S. 1965. *The genesis of modern management*. Harmondsworth: Penguin.

Samuel, R. 1977. Workshop of the world: Steam power and hand technology in mid-Victorian Britain. *History Workshop Journal* 3(Spring): 6–72.
Schor, J., and S. Bowles. 1984. *The cost of job loss and the incidence of strikes*. Harvard Institute of Economic Research discussion paper no. 1105.
Slichter, S. 1929. The current labor policies of American industries. *Quarterly Journal of Economics* 43(May): 393–435.
Sobel, R. 1983. *IBM: Colossus in transition*. New York: Bantam Books.
Thompson, E.P. 1967. Time, work-discipline, and industrial capitalism. *Past and Present* 38(December): 56–97.
Wilkinson, F. (ed.). 1981. *The Dynamics of labour market segmentation*. London: Academic.
Wood, S. (ed.). 1982. *The Degradation of work?: Skill, deskilling, and the labour process*. London: Hutchinson.
Zimbalist, A. 1975. The limits of work humanization. *Review of Radical Political Economics* 7(2): 50–59.

Labour Supply

Richard Blundell and Thomas MaCurdy

Abstract

The analysis of labour supply is placed in a general framework within which empirical models and their resulting elasticity estimates can be interpreted. An explicitly intertemporal life-cycle structure is developed for the choice of hours and participation. The relationship between economic substitution effects found in the labour supply literature and wage impacts on different concepts of employment is considered. We provide a separate discussion of the main issues surrounding the analysis of family labour supply and the analysis of the impact of taxation. We conclude with a discussion on the interpretation of labour supply elasticities for policy analysis.

Keywords

Benefit take-up; Collective models of the household; Cost functions; Dynamic programming; Employment; Engel curve; Euler equations; Frisch specification; Hicksian effect; Hours worked; Indirect utility function; Labour supply; Linear expenditure system; Marshallian effect; Optimal taxation; Reservation wage; Retirement; Slutsky effect; Tax credits

JEL Classifications

J2

The formal analysis of labour supply in economic research extends back to the 1960s, in the work of Becker (1965), Cain (1966), Hanoch (1965) and Mincer (1960), among others. It was developed further in the 1970s, most importantly in the work of Ashenfelter and Heckman (1974), Burtless and Hausman (1978), Gronau (1974) and Heckman (1974a). It would seem reasonable to ask why interest continues in the study of labour supply and what unanswered questions and puzzles remain.

Policy interest in labour supply continually motivates research on all aspects of the subject. One area of active inquiry evaluates the consequences of the new ideas in tax and welfare reform, especially those related to the growing focus on work requirements in the design of welfare reform and on the supply of effort by top-rate tax payers. Another important topic concerns the impacts of reforms of pension and health-care systems on labour supply decisions in later life. Yet another involves gender inequality and the role of female labour supply in removing gender earnings differences and in supporting family incomes. If in addition to these policy motivations, understanding hours-of-work behaviour lies at the heart of explaining the reasons underlying a variety of key trends in the economy. One is the unprecedented growth in female labour supply across many developed economies since the 1970s; a second is the decline in labour supply among older men over the same period, again a phenomenon common to many developed economies; and a third is the labour supply impact of the growth in the disparity between the labour market returns of the educated and those with little formal training. Add to these questions the importance of labour supply in understanding employment over

the business cycle and over the life cycle, and it becomes clear why labour supply has maintained a prominent position in economic research.

Having established its importance, what does the study of labour supply involve? Although the parameter(s) of interest in a labour supply model may seem obvious, on closer inspection it is not so clear-cut. We are typically interested in examining the reaction of labour supply to a change in the wage. But what measure of labour supply and what measure of the wage? Is it employment – the extensive margin – or hours of work for workers – the intensive margin – that is of key interest? Is it the impact that of an anticipated change in the wage or an unanticipated change in the wage? Are we simply concerned with individual labour supply or does family labour supply matter too?

What labour supply elasticity should be used? The wealth of empirical studies on labour supply has produced a plethora of estimated elasticities and response parameters. Differences between estimates can often be attributed to data measurement issues but, as documented in Blundell and MaCurdy (1999), more often than not, a large component of the differences can be explained by the economic framework within which each the estimates is derived. Apart from hourly wages and other income, are controls for lifetime wages included? What about expected changes in other income sources? The precise conditioning variables included in a labour supply model critically change the interpretation and therefore the comparability of estimated elasticities and response coefficients. It is also clear that labour supply responses differ according to the extensive or intensive margin, especially for women. To understand differences across these margins, the specification of effective budget constraints and the nature of fixed costs matter. For men it may well be that the retirement margin could be a margin of growing importance.

An important role of a review of this type is to provide a coherent framework within which different labour supply models can be compared. It is clearly useful to have an explicitly intertemporal framework, although, as we shall see, perfectly interpretable estimates of some important parameters of interest can be recovered from models that look essentially static. Much of the difference across empirical models reflects differences in data availability, and this provides another argument for this approach. The precise form of income, hours or wage variables available will vary wildly across data sources, but this does not necessarily imply incomparable results. Some data provides longitudinal information on individual wages and hours; other data is repeated cross-section but may have more detailed information on asset or consumption levels.

To set the scene we start with a brief discussion of the standard 'static' labour supply model. We then go on to ask what is meant by employment and how one translates estimates of economic substitution effects found in the labour supply literature into wage impacts relevant for the employment concept. Next we look at the extension to a life-cycle setting. The objective is always to present a framework within which empirical models and their resulting elasticity estimates can be interpreted. We provide a separate discussion of the main issues surrounding the analysis of family labour supply and for analysing the impact of taxation and welfare reform. If the literature in respect to all of these topics is too rich to include all of the key references in the text, but a list of some of the leading references is provided at the end of this review. The review ends with a discussion on the interpretation of labour supply elasticities for policy analysis.

Setting the Scene

In the standard labour supply model as applied to individual decisions at a point in time, choices are made over consumption and leisure hours. In each period of time t each individual i, defined by characteristics ν_{it}, has preferences over consumption and leisure hours described by a (within) period utility

$$U(c_{it},\ l_{it};\nu_{it}) \tag{1}$$

in which c_{it} and l_{it} are within-period consumption and leisure hours respectively. (The important

extension to family labour supply is considered below.) The elements of the vector ν_{it} alter preferences, both through observed characteristics of the individual and through this person's unobserved factors influencing 'tastes'. This utility is assumed to be maximized subject to the budget constraint

$$c_{it} + w_{it}l_{it} = y_{it} + w_{it}T \tag{2}$$

in which w_{it} is the hourly wage rate, y_{it}, is non-labour income and T is the total time available for work and leisure.

Non-labour income is made up of two components: asset income and other unearned income. Assuming beginning of period assets, denoted A_{it}, earn a return r_{it} during period t, the former is $r_{it}A_{it-1} + \Delta A_{it}$, in which ΔA_{it} denotes capital gains. Other unearned income is primarily benefit or transfer income and denoted g_{it}. The r.h.s. of (2) is often defined as 'full income' and we denote this income concept as m_{it} throughout, so that

$$m_{it} = y_{it} + w_{it}T. \tag{3}$$

First-order conditions take the familiar form:

$$U_c(c_{it},\ l_{it};\ \nu_{it}) = \lambda_{it} \tag{4}$$

and

$$U_l(c_{it},\ l_{it};\ \nu_{it}) \geq \lambda_{it}w_{it} \tag{5}$$

where λ_{it} is the marginal utility of income. The inequality in (5) determines the reservation wage rule for labour market participation.

Solving for λ_{it} using the budget constraint (2) yields the (Marshallian) decision rule and

$$l_{it} = l(w_{it},\ m_{it};\ \nu_{it}) \leq T_{it} \tag{6}$$

where m_{it} is full-income defined in (3) above. Equivalently we have the hours of work rule

$$h_{it}\begin{cases} = h^s(w_{it}, y_{it}; \nu_{it}) \ \ \forall\{w_{it}, y_{it}; \nu_{it}\} \\ \ni U_L(c_{it}, l_{it}; \nu_{it}) \geq w_{it}\, U_c(c_{it}, l_{it}; \nu_{it}) \\ = 0 \ \ otherwise \end{cases} \tag{7}$$

where y_{it} is defined as in (2).

Preferences over hours of work can, of course, be written analogously to direct utility (1) as

$$U(y_{it},\ T - h_{it};\ \nu_{it}); \tag{8}$$

or by the expenditure (that is, cost) function

$$\xi_{it} = \xi(w_{it},\ V_{it};\ \nu_{it}); \tag{9}$$

or by the indirect utility function

$$V_{it} = V(w_{it},\ y_{it};\ \nu_{it}). \tag{10}$$

The expenditure function solves the problem

$$\begin{aligned} \xi_{it} &= \xi(w_{it},\ V_{it};\ \nu_{it}) \\ &= \min\ c_{it} + w_{it}(T - h_{it})\ \ subject\ to\ V \\ &= U(y_{it},\ T - h_{it};\ \nu_{it},\ \theta); \end{aligned} \tag{11}$$

and the indirect utility inverts the expenditure function to obtain a solution for V_{it}. Whether analysis is conducted with the direct utility, expenditure function, indirect utility or the labour supply equation will depend largely on the approach to estimation.

The inequality (7) represents a corner solution for hours of work and can be stated as a reservation wage condition for participation $w_{it} \geq w^*_{it}$, where w^*_{it} is derived by inverting $h^s(w_{it},\ y_{it};\ \nu_{it}) = 0$. The key econometric problem that follows from this corner solution is that w will not be observed when $h = 0$. Consequently a specification for wages is also required and together they create the *selection problem* addressed by Gronau (1974) and Heckman (1974a, 1979).

Substitution and Income Effects

In a static framework the literature typically cites two types of substitution effects when describing how labour supply responds to changes in the wage rate. First, the uncompensated (or Marshallian) effect refers to the following derivative of labour supply function (7):

$$\frac{\partial h^s}{\partial w} \tag{12}$$

which holds non-labour income y_{it} constant when measuring how much hours of work respond to a shift in wages. If second, one can derive an expression for the compensated labour supply function by computing the derivative of the expenditure function ξ_{it} with respect to w_{it}, and then constructing a function defined as T minus this derivative. This compensated function holds utility constant, and its derivative with respect to w_{it} measures the compensated (or Slutsky or Hicksian) effect. A familiar relationship linking compensated and uncompensated substitution effects is the Slutsky decomposition given by:

$$\frac{\partial h^s}{\partial \omega}|_u = \frac{\partial h^s}{\partial w} + h\frac{\partial h^s}{\partial y}, \tag{13}$$

where the derivative $\frac{\partial h^s}{\partial y}$ shows the impact of changing income on hours of work holding wages constant.

Regular integrability conditions from optimization theory imply that the compensated substitution effect is non-negative

$$\frac{\partial h^s}{\partial \omega}|_u \geq 0. \tag{14}$$

In sharp contrast, the compensated effect $\frac{\partial h^s}{\partial w}$ can be negative or positive depending on the strength of the income effect on labour supply. When $\frac{\partial h^s}{\partial w}$ is negative labour supply is said to be 'backward bending'.

Empirical Evidence

The empirical analysis of the standard labour supply model described here tends to distinguish individuals by gender and by whether there are children at home, finding rather different elasticities across these groups (see Johnson and Pencavel 1984). Allowing for a separate impact of the way the market wage affects the employment and the hours decision has proven to be essential. This partly reflects fixed costs of work and the workings of the welfare system, to be discussed below, but it also highlights the strong evidence that labour supply responses at the extensive margin dominate those at the intensive margin; see Blundell and MaCurdy (1999) for a review of this evidence.

Some Popular Labour Supply Specifications

In discussing particular specifications it is useful to be able to move between all three representations of preferences over labour supply (8)–(10). For example, if the focus is on taxation and welfare participation it is typical to express decisions as a multinomial choice problem over discrete hours choices and work with the direct utility specification. This will be discussed below.

To complete this brief review of the standard labour supply model we consider four popular specifications. The *linear expenditure system* assumes the direct utility function

$$\begin{aligned} U(c_{it},\ T-h_{it};\ v_{it}) = \ & \beta_h(v_{it})\ \ln[T-h_{it}-\gamma_h(v_{it})] \\ & +\beta_c(v_{it})\ \ln[c_{it}-\gamma_c(v_{it})], \end{aligned} \tag{15}$$

where the notation $\beta_h(v_{it})$, $\beta_c(v_{it})$, $\gamma_h(v_{it})$ and $\gamma_c(v_{it})$ indicates that the preference parameters β_h, β_c, γ_h and γ_c are functions of individual attributes v_{it} and therefore can vary across members of the population. (Imposing the restriction $\beta_h(v_{it}) + \beta_c(v_{it}) = 1$ identifies these coefficients.) Abstracting from the dependence on heterogeneous tastes v_{it}, the expenditure function (9) implied for the *linear expenditure system* takes the form:

$$\xi(w,\ V) = \gamma_h w + \gamma_c + w^{\beta_h}\ V;$$

and the uncompensated labour supply function is:

$$h^s(w,\ y) = T - \gamma_h - \frac{\beta_h}{w}(m - \gamma_h w - \gamma_c). \tag{16}$$

A second popular preference specification is the *linear labour supply*

$$h = \alpha + \beta w + \gamma y \tag{17}$$

(for example, see Hausman 1981, 1985a), which comes from the indirect utility function:

L

$$V(w,\ y) = e^{\gamma w}\left(y + \frac{\beta}{\gamma}w - \frac{\beta}{\gamma^2} + \frac{\alpha}{\gamma}\right) \text{ with } \gamma \le 0 \text{ and } \beta \ge 0. \tag{18}$$

Note that since $\partial h/\partial y = \gamma > 0$, the Slutsky condition (13) all but requires $\beta > 0$, ruling out backward bending labour supply. It is arguable that this linear specification allows too little curvature with wages.

Alternative *semilog specifications* and their generalizations are also popular in empirical work. For example, the semilog specification

$$h = \alpha + \beta \ \ln\ w + \gamma y \tag{19}$$

with indirect utility

$$V(w,\ y) = \frac{e^{\gamma w}}{\gamma}(\alpha + \beta \ \ln\ w + \gamma y) + \frac{\beta}{\gamma}\int_{-\gamma w}\frac{e^{-t}}{t}dt \text{ with } \gamma \le 0 \text{ and } \beta \ge 0. \tag{20}$$

Moreover, the linearity of (19) in α and $\ln w$ makes it particularly amenable to an empirical analysis with unobserved heterogeneity, endogenous wages and non-participation as discussed below (see Blundell et al. 1998).

Neither (17) nor (19) allows backward bending labour supply behaviour, although it is easy to generalize (19) by including a quadratic term in $\ln w$. Note that imposing integrability conditions at zero hours for either (17) or (19) implies positive wage and negative income parameters. A simple specification that does allow backward bending behaviour, while retaining a three parameter linear in variables form, is that used in Blundell et al. (1992):

$$h = \alpha + \beta \ \ln\ w + \gamma\frac{y}{w} \tag{21}$$

with indirect utility

$$V(w,\ y) = \frac{w^{1+\gamma}}{1+\gamma}\left(\alpha - \frac{\beta}{1+\gamma} + \beta \ \ln\ w + (1+\gamma)\frac{\gamma}{w}\right) \text{ with } \gamma \le 0 \text{ and } \beta \ge 0; \tag{22}$$

see Stern (1986). This form has similar properties to the specification of Heckman (1974a, b, c). Further empirical specifications are described in Blundell et al. (2007), where the econometric issues of dealing with the extensive margin and missing wages are discussed in detail.

The Impact of Wages and Income on Hours of Work and Employment

Addressing many of the questions asked by policymakers about labour supply involves evaluating the extent to which employment in a population can be expected to change in response to a shift in the returns to work. Relying on existing empirical work to answer such questions requires resolution of two issues: (1) what is meant by employment?; and (2) how does one translate estimates of economic substitution effects found in the labour supply literature into wage impacts relevant for the relevant employment concept?

Three Concepts of Employment and Labour Supply

There are three distinct concepts of labour supply or expected hours of work, which are often confused in the literature. Consider a population of consumers all of whom receive a common wage w and non-labour income y, but who have different tastes v_{it}'s. Let the density function $f(v)$ denote the distribution of 'preferences for work' over the population.

One measure of labour supply is the fraction of the population who works:

$$P(w,y) = \Pr(h^s(w,\ y;\ v_{it}) > 0) = \int_{\Theta} f(v)dv \text{ where } \Theta = \{v_{it} : h^s(w,\ y;\ v_{it}) > 0\}. \tag{23}$$

A second concept is the average hours worked among those employed:

$$\frac{E\left(h^s(w_{it},\ y_{it}; v_{it}) \mid h^s_{it} > 0\right) = \int_{\Theta} h^s(w,\ y;\ v_{it})f(v)\ dv}{P(w,\ y).} \tag{24}$$

Yet a third measure of labour supply is the average hours worked in the entire population:

$$E(h^s(w_{it},\ y_{it}; v_{it})) = \int_{\Theta} h^s(w,\ y;\ v_{it}) f(v)\ dv. \tag{25}$$

While these three measures of labour supply depend on many of the same parameters, they are clearly distinct concepts. If a researcher is interested in the effect of wages on employment, then the derivative of (23) with respect to w measures the appropriate quantity. If, instead, one wants to know how much an increase in the wage rate affects total aggregate hours of work, then the derivative of (25) with respect to w gives the relevant measure.

There is also some confusion in the literature concerning the appropriate interpretation of the partial derivatives of these different measures of labour supply. The partial derivatives of the hours of work function given by (7), h^s_w and h^s_y, produce the textbook uncompensated wage and income effects. Casual inspection of (23) reveals that the derivatives of $P(w, y)$ with respect to w and y do not correspond to h^s_w and h^s_y (Lewis 1967; Ben-Porath 1973). Whereas P_w must be positive, h^s_w need not be. Moreover, the partial derivatives of (24) or (25) with respect to w and y do not correspond to the uncompensated substitution and income effects, h^s_w and h^s_y, unless the inequality condition (7) is satisfied for everyone in the population and the labour supply function h^s takes a special form. These simple points have been ignored in much of the literature. For example, Hall (1973) and Boskin (1973) interpret the partial derivative of estimates of Eq. (25) with respect to w and y as estimates of h^s_w and h^s_y respectively. Others interpret partial derivatives of (24) (estimated from labour supply functions fit on samples of working individuals) as estimates of the Marshallian–Hicks–Slutsky parameters. If non-participation is a significant phenomenon in the population being sampled, estimates of (23), (24) nor (25) do not generate meaningful structural labour supply parameters.

Aggregate Labour Supply

Conditions have been established for utility functions that enable one to aggregate micro labour supply functions to obtain economically meaningful market functions. Satisfaction of these conditions implies equivalency of micro and macro substitution effects. In the case when consumers face a common set of prices and have different incomes, Gorman's (1961; 1976) seminal contributions specify those sets of preference consistent with linear Engel curves, which he shows are required properties of preference to carry out exact aggregation of micro demand functions to macro formulations. The macro specification is a 'representative consumer' version of the original individual preference relationship. Gorman's conditions are insufficient for aggregation of labour supply functions since wages, in contrast to prices, vary considerably across individuals in any interesting empirical application. Muellbauer (1981) refines Gorman's aggregation conditions to apply to the labour supply case allowing for wages along with income to different across individuals.

For a market labour supply function to have a form consistent with the underlying micro specifications aggregated to derive its construction, the expenditure function (9) must necessarily take the general form:

$$\xi(w_{it},\ V_{it};\ v_{it}) = \alpha_t(v_{it}) + w_{it}\beta_t + w^{\delta}_{it} b_t V_{it}. \tag{26}$$

(Inspection of the specification – the equation above Eq. (16) – for $\xi(w_{it},\ V_{it};\ v_{it})$ for the *linear expenditure system* reveals that it has the form required by (26) when $\beta_h(v_{it.}) = \beta_h$, $\beta_c(v_{it}) = \beta_c$, and $\gamma_h(v_{it}) = \gamma_h\ \ \forall\ \ v_{it}$) The uncompensated labour supply function implied by (26) is given by:

$$h^s(w_{it},\ y_{it};\ v_{it}) = \pi_t - \frac{\delta}{w_{it}}(y_{it} - \alpha_t(v_{it})) \tag{27}$$

where

$$\pi_t = (1-\delta)(T - \beta_t). \tag{28}$$

L

In this specification, only the preference components $\alpha(v_{it})$ can vary across individuals in the static setting. Rather than expressing this relationship as hours of work, one typically finds (9) it written as the earning function:

$$w_{it}h_{it}^{s} = \pi_t w_{it} - \delta y_{it} + \delta\alpha_t(v_{it}). \qquad (29)$$

Given its linear structure, one clearly sees that estimation of the micro and aggregate substitution and income effects corresponds to the same preference parameters. Viewed in a pooled cross-section time-series context, the preference components α_t, β_t, and b_t typically will be functions of prices in period t which are common across individuals in the cross section corresponding to the period, but these prices do change over time. To create a valid form for preferences, the α_t and β_t must be homogeneous of degree 1 in prices, and b_t must be homogeneous of degree zero.

What concept of labour supply does this aggregate relationship represent? In a world where everyone works, the average of (27) corresponds to both the expected values of hours worked among the employed (24) and overall populations (25); after all, these are exactly the same samples. Moreover, the economic concept of the uncompensated substitution effect directly measures the response one would estimate using an empirical specification based on either Eq. (24) or Eq. (25).

These nice relationships, however, entirely break down when one recognizes that the employment decision is typically influenced by a change in wages, be it across people or a shift in the distribution that occurs over time. With the no-work/work decision being affected for some people, impacts now critically depend on the properties of distribution of preferences determined by the density function $f(v)$, which could itself shift over time. The effects of wages on the three concepts of labour supply given by (23), (24) and (25) again become distinct, and none directly measures the economic notions of substitution effects outlined above. When labour market participation is a choice in the population, *no* conditions exist for consistently aggregating micro labour supply function to obtain a macro function that can be given a coherent 'representative agent' interpretation. Substitution effects estimated in an aggregate setting cannot be interpreted coming from a single agent-optimizing framework, and the wage effects estimated from micro data considered alone will typically provide insufficient information to project aggregated impacts.

Labour Supply Over the Life Cycle

Although its study is often placed in an effectively static framework as in (1) and (2), labour supply is clearly part of a lifetime decision-making process. Individuals attend school early in life, accumulate wealth while in the labour force, and make retirement decisions late in life; each of these activities can only be understood in a life-cycle framework. We know that savings from labour earnings are often required to sustain individuals, or their dependants, during periods when they are out of the labour market. In addition, variations in health status, family composition and real wages provide incentives for individuals to vary the timing of their labour market earnings for income-smoothing and insurance purposes.

To keep things simple we assume life-cycle utility at time t has the form

$$\mathscr{U}_{is} = E_t\left\{\sum_{t=s}^{L}\frac{1}{1+\delta_t}U(c_{it}, l_{it}; v_{it})\right\} \qquad (30)$$

in which E_t is the expectations operator conditional on information up to and including period t and where δ_t is the subjective discount rate. Maximization of (30) takes place subject to at intertemporal budget constraint. For this we need to write down the path of assets:

$$A_{it+1} = A_{it} + r_t A_{it} + b_{it} + w_{it}h_{it} - c_{it} \qquad (31)$$

where A_{it} is the assets held at the beginning of period t and $r_{t.}$ is the return on assets earned in period t.

The form of life-cycle preferences and of the budget constraint in (30) and (31) is not innocuous. The time-separability of (30) rules out habits

and slow adjustment. The rA term in (31) assumes that individuals can borrow and lend via the simple credit market at rate r and consequently rules out borrowing constraints.

Nevertheless, under these assumptions the first-order conditions (4) and (5) continue to hold and to determine within-period allocations of time and consumption. Intertemporal allocations are determined through the choice of the marginal utility of consumption λ_t in (4). Consequently allocations over the life cycle will be summarized through the evolution of λ_t.

To understand these conditions in an intertemporal context we can use the knowledge that λ_{it}, the marginal utility of wealth, evolves over time according to

$$\lambda_{it} = \frac{1}{1+\delta_t} E_t\{\lambda_{it+1}(1+r_{it})\} \qquad (32)$$

where the real interest rate r_{it} is allowed to be stochastic. Relationship (32) is often referred to as the stochastic Euler equation (see Hansen and Singleton 1983).

Frisch (λ-constant) Labour Supply Equations

Frisch, or marginal-utility-of-wealth 'λ' constant, labour supply functions provide an extremely useful method for analysing life-cycle maximization problems (see Browning et al. 1985). In this framework, the marginal utility of wealth, λ, serves as the sufficient statistic which captures all information from other periods that is needed to solve the current-period maximization problem. The time-separable form of the utility maximizing model implies that the marginal within-period decisions depend on the past and future through the single 'sufficient statistic' λ_{it}. Even though the marginal utility of wealth λ_{it} is not observable to the empirical economist, the rule for its evolution (32) enables a method of moments estimation of the labour supply parameters.

To briefly see how estimation takes place in this framework, consider the simple parametric form for preferences chosen in MaCurdy (1981). The utility specification MaCurdy used does not allow for corner solutions and takes the form

$$U_t = \theta_t c_t^{\gamma} - \varphi_t h_t^{\alpha} \; 0 < \gamma < 1, \;\; \alpha > 1 \qquad (33)$$

where $ht.$ corresponds to hours of work and ct to consumption. The range of parameters ensures positive marginal utility of consumption, negative marginal utility of hours of work and concavity in both arguments. The Frisch labour supply is

$$\log h_t = \theta_t^* + \log \lambda + \frac{1}{\alpha - 1} \ln w_t + \frac{\rho - r}{\alpha - 1} t \qquad (34)$$

where the use of log hours of work presumes that all individuals work and hence $h > 0$. In (34) λ is the shadow value of the lifetime budget constraint and t is the age of the individual. Finally A_t^* reflects preferences and is defined by $\theta_t^* = -\frac{1}{\alpha-1} \log \; \theta_t$. This equation has a simple message: Hours of work are higher at the points of the life cycle when wages are high $\left(\frac{1}{\alpha-1} > 0\right)$. Moreover if the personal discount rate is lower than the interest rate, hours of work decline over the life cycle. Finally, hours of work will vary over the life cycle with θ_t^*, which could be a function of demographic composition or other taste shifter variables.

The MaCurdy (1981) paper set out the first analysis of issues to do with estimating intertemporal labour supply relationships. However the approach did not deal with corner solutions and the extensive margin, which is particularly relevant for women. The first attempt to do so, in the context of a life-cycle model of labour supply and consumption is the paper by Heckman and MaCurdy (1980). In this model women are endowed with an explicitly additive utility function for leisure l and consumption c in period t, of the form:

$$U_t = \theta_t \frac{l_t^{\alpha} - 1}{\alpha} + \varphi_t \frac{c_t^{\gamma} - 1}{\gamma} \;\; \alpha, \;\; \gamma < 1. \qquad (35)$$

Optimization is assumed to take place under perfect foresight. Solving for the first-order conditions we obtain the following equation for leisure

$$\ln l_t \begin{cases} = \theta_t^* \frac{1}{\alpha - 1} \ln w_t + \frac{\rho - r}{\alpha - 1} t + \lambda^* \; \textit{when the woman works} \\ = \ln \bar{l} \;\; \textit{otherwise} \end{cases} \qquad (36)$$

L

where

$$\lambda^* = \frac{1}{\alpha - 1} \ln \lambda \text{ and } \theta_t^* = -\frac{1}{\alpha - 1} \ln \theta_t. \quad (37)$$

Two-Stage Budgeting and Marshallian Labour Supply Equations

In this time-separable optimizing problem there are alternative 'sufficient statistics' to the marginal utility of wealth that completely summarize the past and future as it impacts on the period t labour supply decision. From Gorman (1959, 1968), intertemporal separability implies that the decision rule can be thought of in two stages. First allocate to period t according to

$$m_{it} = M(w_{it}, y_{it}, A_{it-1}, r_t, v_{it}, z_{it}) \quad (38)$$

where z_{it} represents the information used to form expectations of future real wages and other household attributes that are uncertain at time t. At the second stage, given m_{it}, the within-period first-order conditions (4) and (6) remain valid. Moreover, the estimation of 'm-conditional' labour supply functions are robust to liquidity constraints and other capital market imperfections.

Marginal Rate of Substitution Equations

Eliminating λ_{it} from the first-order conditions (4) and (6) yields the marginal rate of substitution function

$$MRS_l(c_{it}, l_{it}; v_{it}) \geq w_{it} \quad (39)$$

where

$$MRS_l(c_{it}, l_{it}; v_{it}) = \frac{U_l}{U_c}. \quad (40)$$

Again, (39) is robust to liquidity constraints and other capital market imperfections. As we know from our general discussion of elasticities, the constant marginal utility of wealth (Frisch) elasticity is greater than the Slutsky-compensated (within-period) elasticity which is again greater than the standard uncompensated Marshallian elasticity, see Blundell (1998).

Relationships Among the Life-Cycle Elasticities

The Frisch specification treats the individual marginal utility of wealth as a 'fixed effect' and allows the researcher to estimate only the intertemporal substitution elasticity. Given that appropriate methods are employed to account for the fixed effect (generally first differencing in panel data), the relevant independent variables, apart from the wage, are simply within-period characteristics and age. The Frisch elasticity, by ignoring this (unexpected) shift in wealth from a once-and-for-all change in real wages, is larger than the policy-relevant elasticity and overestimates the impact of a reform.

Direct estimation of the simple parameterization of the full life-cycle model, required to recover policy-relevant elasticity, relies on specifications for both within-period utility and the individual marginal-utility-of-wealth effect. As a result, controls are needed for all of the following: 'start of life' characteristics, current-period characteristics which affect the within-period utility function, age, expected wages and initial wealth. Expected wages are typically unobservable and initial wealth is generally not included in datasets, so these should be replaced with the parameters governing the time path of wages and property income, which must be jointly estimated with the labour supply equation. Estimation of this full framework allows computation of both the intertemporal substitution elasticity and the elasticity of labour supply in reaction to a full, parametric wage profile shift. However, it is also the most demanding in terms of data.

It is worth noting that the elasticity derived from the static specification which uses unearned income to compute virtual income can be placed in an intertemporal setting but is economically meaningful only under a strong assumption of either complete myopia or perfectly constrained capital markets. Otherwise, this elasticity confuses movements along wage profiles with shifts of these profiles and, thus, yields response parameters which are a mixture of these. Such hybrid estimates lack an economic interpretation and are not generally useful in policy evaluation.

To illustrate the challenges encountered with inferring the different substitution effects from one another, consider a life-cycle extension of the *linear expenditure system* (LES) in a deterministic setting. A multi-period expansion of the static LES utility function given by (15) takes the form:

$$\begin{aligned}\mathscr{U} &= \sum_{t=1}^{\tau} \varphi_t \cdot U(c_{it},\ l_{it};\ v_{it}) \\ &= \sum_{t=1}^{\tau} \varphi_t [\beta_h \ln(T - h_t - \gamma_h) + \beta_c \ln(c_{it} - \gamma_c)],\end{aligned} \tag{41}$$

where the normalization $\sum_{t=1}^{\tau} \varphi_t = 1$ (in addition to $\beta_h + \beta_c = 1$) identifies preference parameters. The specification implied for the life-cycle uncompensated labour supply function for hours of work in period t is:

$$\begin{aligned}h_t^s(\omega,\ R,\ M;\ v) &= T - \gamma_h \\ &\quad - \frac{\varphi_t \beta_h}{\omega_t}\left(M - \sum_{k=1}^{\tau} \gamma_h \omega_k - \sum_{k=1}^{\tau} \gamma_c R_k\right)\end{aligned} \tag{42}$$

where the quantities ω_t denote the discounted value of the period-t wage rate; R_t represents the discounted price of consumption in period t; and M designates the 'full income' equivalent of the individual's wealth. The period-t marginal-utility-of-wealth 'λ' constant labour supply function takes the form:

$$h_{it} = h^{\lambda}(\omega_{it},\ R_{it},\ \lambda_{it}) = T - \gamma_h + \frac{\varphi_t \beta_h}{\lambda_{it} \omega_{it}}. \tag{43}$$

Accordingly, the uncompensated substitution effect associated with a change in wage rate ω_t on hours of work h_t is given by:

$$\begin{aligned}\frac{\partial h_t^s}{\partial \omega_t} &= \frac{\varphi_t \beta_h}{\omega_t^2}\left(y - \sum_{k=1}^{\tau} \gamma_h \omega_k - \sum_{k=1}^{\tau} \gamma_c R_k + \gamma_h \omega_k\right) \\ &= \frac{(T - \gamma_h)(1 - \varphi_t \beta_h)}{\omega_{it}} - \frac{h_t}{\omega_t};\end{aligned} \tag{44}$$

and the intertemporal substitution effect corresponding to change in ωt on $ht.$ is:

$$\frac{\partial h_t^{\lambda}}{\partial \omega_t} = -\frac{\varphi_t \beta_h}{\lambda_{it} \omega_{it}^2} = \frac{(T - \gamma_h)}{\omega_{it}} - \frac{h_t}{\omega_t}. \tag{45}$$

The following relationship links these two hour-of-work responses:

$$\frac{\partial h_t^{\lambda}}{\partial \omega_t} = \frac{\partial h_t^s}{\partial \omega_t} + \frac{(T - \gamma_h)\varphi_t \beta_h}{\omega_{it}}. \tag{46}$$

Finally, if one were to estimate an uncompensated substitution effect relying on a two-stage-budgeting variant of a labour supply function based on LES utility function (41), then one would compute values for:

$$\begin{aligned}\frac{\partial h_t^s}{\partial \omega_t} &= \frac{\beta_h}{\omega_t^2}(y - \gamma_h) \\ &= \frac{(T - \gamma_h)(1 - \beta_h)}{\omega_t} - \frac{h_t}{\omega_t}.\end{aligned} \tag{47}$$

While inspection of these expressions not surprisingly reveals that the different substitution effects depend on common preference parameters, it also clearly indicates that one must exercise serious caution when attempting to infer values of one type of elasticity from any of the others. Relationship (46) shows that how one can vary endowments and preferences to change intertemporal substitution effects while not changing the uncompensated response. Of course, the above discussion has already described the additional complications encountered in any attempt to relate these economic notions of substitution effects to concepts of labour supply relevant for market measures of wage impacts on employment and hours of work which are the core concepts required for policy analyses.

Retirement and Pension Incentives

The study of retirement incentives and labour supply has typically focused on the dynamic effects of benefit entitlement that occur in many pension and social security schemes (Hurd and Boskin 1984). This has resulted in the more formal use of dynamic programming tools; see Blau (1994) and Rust and Phelan (1997), for example.

An important area for current research is the incorporation of these incentives into a life-cycle labour supply model.

Family Labour Supply

For the purposes of this discussion we are concerned with a family or household as comprising two working-age individuals, referred to as husband and wife below. These are the decision-making individuals in the family. Families with a single parent are subsumed in the discussion of the regular labour supply model. The central issue then becomes one of the mechanism whereby labour supply decisions are made within the household. Are they taken in a fully coordinated way as if by a single decision maker – the unitary model – or are they the result of some collective bargain – the collective model?

The Unitary Model of Family Labour Supply

Suppose we can take a family or household as being made up of two working-age individuals, referred to as husband and wife below. Children and any other dependants will be included in the vector of observable household characteristics ν_{it}. For such a household, within period utility may be written

$$U_{it} = U\left(c_{it},\ l^h_{it},\ l^w_{it};\ v_{it}\right) \tag{48}$$

and budget constraint

$$c_{it} + w^h_{it} l^h_{it} + w^w_{it} l^w_{it} = m_{it} \tag{49}$$

where w^h_{it} and w^w_{it} refer to the hourly wage of the husband and wife respectively.

The marginal conditions for the λ-constant (Frisch), Marshallian and marginal rate of substitution labour supply equations described in the previous section follow naturally from the first-order conditions

$$U_c\left(c_{it},\ l^h_{it},\ l^w_{it};\ v_{it}\right) = \lambda_{it}, \tag{50}$$

$$U_h\left(c_{it},\ l^h_{it},\ l^w_{it};\ v_{it}\right) \geq \lambda_{it} w^h_{it} \tag{50}$$

and

$$U_w\left(w_{it},\ l^h_{it},\ l^w_{it};\ v_{it}\right) \geq \lambda_{it} w^w_{it} \tag{52}$$

where the subscripts h and w refer to derivatives with respect to the non-market hours of husband and wife respectively. See Ashenfelter and Heckman (1974), Wales and Woodland (1976) and Blundell and Walker (1982), for example.

Notice that there is still only a single marginal utility of wealth λ_{it} and therefore the extension to the life-cycle framework of the previous section is straightforward. There remains only one life-cycle condition (32). Consequently allocations to each individual in this time-separable model satisfy equality of marginal utility of wealth; see Blundell and Walker (1986), for example.

Collective Family Labour Supply

The advantages of the unitary model are well known: it allows the direct utilization of consumer theory, recovering preferences from observed behaviour in an unambiguous way, and provides a coherent intertemporal framework for interpretation of empirical results. An argument against this approach is that it treats individuals in the family as a single decision-maker rather than as if they were a collection of individuals. Although true, this can be weakened through a simple decentralization argument. Suppose we let c^h and l^h refer to the private consumption of the husband and his own leisure time respectively. Defining the private consumption of the wife in the same way, we may write the within-period household utility as

$$U\left(c_{it},\ l^h_{it},\ l^w_{it};\ v_{it}\right) = \tilde{U}\left(F_h\left(c^h_{it},\ l^h_{it};\ v_{it}\right),\ F_w\left(c^w_{it},\ l^w_{it};\ v_{it}\right)\right) \tag{53}$$

where $F_h\left(c^h_{it},\ l^h_{it};\ v_{it}\right)$ is the sub-utility for the husband and $F_w\left(c^w_{it},\ l^w_{it};\ v_{it}\right)$ is the sub-utility of the wife. Family utility has a 'weakly separable' form and decentralization follows: allocations of total household (full) income are made between each household member and then individuals act as if they are making their labour supply and

consumption decisions conditional on this initial-stage outlay. Of course, even if consumption goods are privately consumed, they are typically only measured at the household level – so that the individual consumptions are 'latent' to the economist.

So what is it that collective models offer? They effectively relax the income allocation rule between individuals so that this allocation can depend on relative wages and other variables in a way that reflects the bargaining position of individuals within the family rather than reflecting the symmetry assumption underlying the joint optimizing framework of the traditional approach. Individuals within the family can be altruistic and allocations Pareto efficient, but still the allocation rule can deviate from the optimal rule in the traditional model.

The most lucid statement of this argument can be found in the papers on household labour supply by Chiappori (1988, 1992). He states the family labour supply problem as one of

$$\max \ \theta U^h + (1-\theta)U^w s.t. \ c_{it} + w_{it}^h l_{it}^h + w_{it}^w l_{it}^w (= m_{it}) = \left(w_{it}^h + w_{it}^w\right)T + y$$

with some non-negative function $\theta = f\left(w_{it}^h,\ w_{it}^w,\ x_{it},\ m_{it}\right)$ representing the weight given to utility U^h. What Chiappori shows is that this is equivalent to a sharing rule solution in which U^h gets income $\phi\left(w_{it}^h,\ w_{it}^w,\ x_{it},\ m_{it}\right)$ out of y, and then allocates according to the rule:

$$\max \ U^h \ s.t. \ c_{it}^h + w_{it}^h l_{it}^h = w_{it}^h T + \phi\left(w_{it}^w,\ w_{it}^w,\ x_{it},\ m_{it}\right)$$

where x_{it} may be a distribution factor.

Conditions for the identification of preferences and the sharing rule (up to a linear translation) simply require an observable private good – here assumed to be the individual's leisure. The intuition behind identification is simple: under the exclusive good assumption the spouse's wage can only have an effect through the sharing rule. Variation of income and wage will then provide an estimate of the marginal rate of substitution in the sharing rule. The same can be done for both spouses, and since the sharing rule must sum to 1, the partial derivatives of the sharing rule can be recovered.

The empirical implementation of the collective model has been slow but is growing in recent years; see Donni (2003) and Fortin and Lacroix (1997), for example. Generalizing the collective model to allow for non-participation and corner solutions requires additional care (see Blundell et al. 2006). The generalization to an intertemporal framework is still in its infancy.

The collective approach is not the only way to conceive of bargaining in family labour supply; see Kooreman and Kapteyn (1990), Lundberg (1988) and McElroy (1981) for important alternatives.

Labour Supply with Taxation and Welfare Participation

The tax and welfare system leads to well documented nonlinearities and non-convexities in the budget constraint facing any individual. This considerably complicates the labour supply problem and, even in the static setting, discrete choice programming methods are required. The basic nonlinear budget constraint problem has been described in detail in Hausman (1985a), Moffitt (1986), MaCurdy et al. (1990) among others.

To further address the issues encountered with nonlinear budget sets, there has been a steady expansion in the use of sophisticated statistical models characterizing distributions of discrete-continuous variables that jointly describe both interior choices and corner solutions in demand systems. These models offer a natural framework for capturing irregularities in budget constraints, including those induced by the institutional features of tax and welfare programmes. Typically the overall stochastic specification is represented by a mixed-multinomial specification across discrete choices over ranges of hours, for example in the work of Hoynes (1996) and Keane and Moffitt (1998). In this research, individuals are assumed to maximize their (stochastic) utility subject to a

budget constraint, determined by a fixed hourly wage and the tax and benefit system. The utility function (8) is often approximated with a second-degree polynomial in hours of work and net income. A common feature of these models is the introduction of unobserved preference heterogeneity in the marginal rate of substitution between work and consumption. Further unobserved heterogeneity in the 'costs' of programme participation and in fixed costs of work is also now commonplace; see Blundell and MaCurdy (1999).

Discrete Hours Choices

In view of the large number of non-convexities, it is common to discretize hours into hours bands, and consider the choice across these intervals. For example, in Keane and Moffitt (1998) the utility function is modelled as

$$U^*_{H^j} = U\left(y_{H^j},\ T - H^j; x\right) + \varepsilon H^j \qquad (54)$$

where εH^j represents an unobserved preference component relating to the particular hours choice $h \equiv H^j$, assumed to be distributed as an extreme value random variable. Household disposable income, when supplying H^j hours, is defined by

$$y_{H^j} = wH^j + b - R\left(H^j,\ w,\ g; x\right) \qquad (55)$$

where w is the pre-tax hourly wage rate, g is other income (not including benefits and transfers) and $R(H^j,\ w,\ g; x)$ is the tax payable (positive or negative) when working H^j hours and having demographic composition x. Thus R will reflect both tax payments and credits or welfare payments received. This expression reflects the fact that the tax and benefit system may be nonlinear and may give rise to non-convexities; in these cases it is no longer possible to express the impact of the tax system simply by a marginal tax rate.

Fixed Costs of Work

Fixed costs are the costs that an individual has to pay to get to work; see Cogan (1980, 1981) and Hausman (1980). For parents, they are made up in part by childcare costs. In particular, childcare induces both fixed and variable costs that effectively act as a marginal tax rate. However, there are additional costs, for example, transport, which will vary by household type and by region. These are typically modelled as a once-off weekly cost and are subtracted directly from net income for any choices that involve work. They enter the utility comparisons in each individual's work–non-work choice.

Missing Wages

For non-workers gross wages are not observed. As in the discussion of corner solutions and non-participation in Section 1, for each individual we could write the logarithm of hourly wages as

$$\ln\ w = z'\gamma + \omega \qquad (56)$$

where ω has density $g(\omega)$ and where z will include education, cohort and time dummies and their interactions. In principle the wage equation and the labour supply model can be estimated jointly. However, for computational reasons it is common to pre-estimate the marginal density of wages and then treat it as known at the estimation stage. This method can account for the endogeneity of gross wages and also allows for the complex relationship between gross wages and marginal wages in the tax and benefit system.

Programme Participation, Stigma and Benefit Take-Up

Since the important work of Moffitt (1983) and Ashenfelter (1983), the formal analysis of welfare stigma and programme participation has been a key component of the labour supply impacts of tax and welfare programmes. Suppose $P = 1$ indicates that an eligible individual participates in a welfare programme. Eligibility at any hours point H^j will typically depend on earnings, other income sources, family characteristics, and the rules of the tax and benefit system. Suppose that the hassle cost and stigma is given by η, an unobservable random variable. Then we may express utility for combination $\{H^j, P\}$ as

$$U^* \equiv U^*\left(y_{H^j,P} - F,\ T - H^j, | x\right) - \eta P \qquad (57)$$

where F is fixed costs of work. The stigma cost variable η may be modelled as a single unknown parameter representing a common cost across all individuals. More usefully it can be modelled as a random process with unknown mean μ_η and distribution $f_\eta(\eta)$. The parameters of its distribution are then recovered during estimation. Notice that net income $Y_{Hj;P}$ also depends directly on P through the working of the benefit and credit system. For any distribution of stigma costs an increase in the generosity of the benefit will increase the probability of take-up. Consequently, other things equal, take-up will be higher among those eligible for a larger benefit.

As documented in Blundell and MaCurdy (1999), for each hours H^j where the family is eligible to participate in the programme, utility function (57) defines a reservation stigma cost $\eta^*_{H^j}$ above which the family would prefer not to participate at that hours level (note that the same family may choose to participate for some other hours level where it is also eligible for the programme). Given the family characteristics and the tax/benefit rules, the eligibility of each family at each level of hours can be determined, and the likelihood used in estimating the unknown parameters of labour supply, wages, fixed costs and programme participation can be fully specified.

Family Labour Supply and Taxation

The modelling structure for couples requires but few modifications provided a 'unitary' model of family labour supply is adopted. The important difference in practice, as far as taxation and welfare is concerned, is that now we have to take into account the interaction of the welfare benefits that individuals may receive; see Hausman and Ruud (1984), Hoynes (1996) and van Soest (1995). Thus, the options facing each spouse are typically very different depending on whether the other family members work. Tax credit systems tend to lead to complex interactions between the effective tax rates for spouses (see Blundell et al. 2000; Eissa and Hoynes 2004).

Optimal Taxation and Labour Supply

One of the key developments in the use of labour supply elasticities has been in the design of 'optimal' tax and transfer systems following the innovative work of Saez (2001, 2002) and Laroque (2004). This has established a close link between the empirical analysis of labour supply responses and the early literature on optimal taxation (Mirrlees 1971); see for example the implementation of these ideas in Immervol et al. (2007).

Randomized Control Trials and Quasi-Experimental Approaches

Focusing purely on the reduced form impact of tax reform on labour supply, there have been several influential studies that have sidestepped the labour supply choice model and attempted to recover the impact of reforms on labour supply using randomized control experiments and quasi-experiments. The leading pure experiments are the Seattle–Denver Income Maintenance Experiment documented in Ashenfelter and Plant (1990) and the more recent Canadian Self Sufficiency Program for single mothers on welfare analysed in Card and Robins (1998). These provide a direct impact of a specific reform and also provide a useful basis from which to judge estimates from structural models.

Quasi-experimental methods, which compare an eligible and a comparison group before and after a reform, have also been influencial – for example the Eissa and Liebman (1996) study of the 1986 expansion of the Earned Income Tax Credit in the United States and the impact of tax rate changes on the taxable earnings of higher-income earners; see, in particular, the study by Feldstein (1995) and the further analysis by Gruber and Saez (2002). However, these quasi-experimental approaches require strong assumptions to be interpretable as measuring behavioural responses; see Blundell and MaCurdy (1999).

Conclusions: Which Labour Supply Elasticities for Policy Evaluation?

An argument has been made for an explicitly intertemporal framework, although, as we have seen, perfectly interpretable estimates of some important parameters of interest can be recovered from models that look essentially static. Much of

the difference across empirical models reflects differences in data availability, and this provides another motivation for our approach. Precisely what form of income, hours or wage variables is available will vary widely across data sources, but this doesn't necessarily imply incomparable results. Some data provides longitudinal information on individual wages and hours; other data is repeated cross section but may have more detailed information on asset or consumption levels.

In whatever context the analysis of labour supply takes place, estimation will benefit from exogenous wage and income variation. One thing is clear: the type of trends that have occurred in many economies since the 1970s and the wide range of policy reforms designed to change labour supply incentives do strengthen the case for exploiting time-series information and avoiding complete reliance on purely cross-section data.

Four basic elasticities have been described which cover the main wage elasticities estimated in empirical labour supply analysis. Two are within-period elasticities: the first relating to the purely static formulation and the second relating to the two-stage budgeting specification. Two are life-cycle elasticities: the first being the intertemporal elasticity of substitution relating to the Frisch specification and measuring responses to evolutionary movements along the life-cycle wage profile, and the second relating to a full life-cycle specification and measuring responses to parametric shifts in the life-cycle profile itself. As most tax and benefit reforms are probably best described as once-and-for-all unanticipated shifts in net-of-tax real wages today and in the future, the most appropriate elasticity for describing responses to this kind of shift is the last of these. For the standard business cycle model it is the anticipated change that is of importance. As we have noted, these two elasticities can be substantially different due to income and wealth effects.

If a researcher regresses log hours of work on age; all age-invariant characteristics determining lifetime wages, preferences, and initial permanent income; and log wage, then the coefficient on the current wage rate is the Frisch elasticity. Intuitively, this approach controls for differences in the initial value of the marginal utility of wealth across consumers and leaves higher-order age variables as instruments to identify wage variation. Hence, only evolutionary wage variation along the age–wage path is included.

If, alternatively, a researcher regresses log hours worked on property income, age, age squared, and log wage, the coefficient on wage is the response of labour supply to a parametric wage shift – including both the intertemporal substitution effect and the reallocation of wealth across periods captured by a change in the marginal utility. Intuitively, this approach controls for age effects and leaves individual characteristics as instruments for wage. Changes in these characteristics capture full profile shifts rather than movements along the age–wage path.

The standard static labour supply representations fit neither of these patterns, as they include property income together with personal characteristics rather than age and age squared. Hence, given the existence of life-cycle effects they confuse the effect of movements along the wage profile with shifts in the profile and, thus, yield parameters without an economic interpretation.

See Also

- ▶ Collective Models of the Household
- ▶ Elasticity of Intertemporal Substitution
- ▶ Hours Worked (Long-Run Trends)
- ▶ Indirect Utility Function
- ▶ Retirement
- ▶ Substitutes and Complements
- ▶ Taxation of Income
- ▶ Taxation of the Family

Bibliography

Abbott, M., and O. Ashenfelter. 1976. Labor supply, commodity demand and the allocation of time. *Review of Economic Studies* 43: 389–411.

Altonji, J.G. 1982. The intertemporal substitution model of labour market fluctuations: An empirical analysis. *Review of Economic Studies* 49: 783–824.

Altonji, J.G. 1986. Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy* 94: S176–S215.

Altug, S., and R. Miller. 1990. Household choices in equilibrium. *Econometrica* 58: 543–570.

Altug, S., and R. Miller. 1998. The effect of work experience on female wages and labour supply. *Review of Economic Studies* 65: 45–85.

Apps, P.F., and R. Rees. 1997. Collective labor supply and household production. *Journal of Political Economy* 105: 178–190.

Arellano, M., and C. Meghir. 1992. Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *Review of Economic Studies* 59: 537–559.

Arrufat, J.L., and A. Zabalza. 1986. Female labour supply with taxation, random preferences, and optimization errors. *Econometrica* 54: 47–63.

Ashenfelter, O. 1983. Determining participation in income-tested social programs. *Journal of the American Statistical Society* 78: 517–525.

Ashenfelter, O., and J. Ham. 1979. Education, unemployment and earnings. *Journal of Political Economy* 87: S99–S166.

Ashenfelter, O., and J.J. Heckman. 1974. The estimation of income and substitution effects in a model of family labor supply. *Econometrica* 42: 73–85.

Ashenfelter, O., and M.W. Plant. 1990. Nonparametric estimates of the labor-supply effects of negative income tax programs. *Journal of Labor Economics* 8: S396–S415.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Ben-Porath, Y. 1973. Economic analysis of fertility in Israel: Point and counterpoint. *Journal of Political Economy* 81: S202–S233.

Bingley, P., and I. Walker. 1997. The labour supply, unemployment and participation of lone mothers in in-work transfer programs. *Economic Journal* 107: 1375–1390.

Blau, D.M. 1994. Labor force dynamics of older men. *Econometrica* 62: 117–156.

Blau, D., and P. Robins. 1988. Child-care costs and family labor supply. *Review of Economics and Statistics* 70: 374–381.

Blomquist, N.S. 1983. The effect of income taxation on the labour supply of married men in Sweden. *Journal of Public Economics* 22: 169–197.

Blomquist, S., and W. Newey. 2002. Nonparametric estimation with nonlinear budget sets. *Econometrica* 70: 2455–2480.

Blundell, R. 1998. Consumer Demand and Intertemporal Allocations: Engel, Slutsky and Frisch, In Stienar Strom (ed), *The Raynar Frisch Centennial Symposium*, Econometric Society Monographs 31, Cambridge University Press, Cambridge.

Blundell, R. 2006. Earned income tax credit policies: Impact and optimality: The Adam Smith lecture. *Labour Economics* 13(4): 423–443.

Blundell, R., and H. Hoynes. 2004. Has in-work benefit reform helped the labour market? In *Seeking a premier league economy*, ed. R. Blundell, D. Card, and R.B. Freeman. Chicago: University of Chicago Press.

Blundell, R.W., and T. MaCurdy. 1999. Labor supply: A review of alternative approaches. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, Vol. 3A. Amsterdam: North-Holland.

Blundell, R.W., and I. Walker. 1982. Modelling the joint determination of household labour supplies and commodity demands. *Economic Journal* 92: 58–74.

Blundell, R.W., and I. Walker. 1986. A life-cycle consistent empirical model of family labour supply using cross-section data. *Review of Economic Studies* 53: 539–558.

Blundell, R.W., J. Ham, and C. Meghir. 1987. Unemployment and female labour supply. *Economic Journal* 97: 44–64.

Blundell, R.W., C. Meghir, E. Symons, and I. Walker. 1988. Labour supply specification and the evaluation of tax reforms. *Journal of Public Economics* 36: 23–52.

Blundell, R.W., A. Duncan, and C. Meghir. 1992. Taxation in empirical labour supply models: Lone mothers in the UK. *Economic Journal* 102: 265–278.

Blundell, R., C. Meghir, and P. Neves. 1993. Labour supply and intertemporal substitution. *Journal of Econometrics* 59: 137–160.

Blundell, R., A. Duncan, and C. Meghir. 1998. Estimating labor supply responses using tax reforms. *Econometrica* 66: 827–861.

Blundell, R., A. Duncan, J. McCrae, and C. Meghir. 2000. The labour market impact of the working families' tax credit. *Fiscal Studies* 21: 75–104.

Blundell, R., P.-A. Chiappori, and C. Meghir. 2005. Collective labour supply with children. *Journal of Political Economy* 113: 1277–1306.

Blundell, R., P.A. Chiappori, T. Magnac, and C. Meghir. 2006. Collective labor supply: Heterogeneity and non-participation. *Review of Economic Studies* 74: 417–445.

Blundell, R.W., T. MaCurdy, and C. Meghir. 2007. Estimation and specification in labor supply models. In *Handbook of econometrics*, ed. J.J. Heckman and E. Leamer, Vol. 7. Amsterdam: North-Holland (forthcoming).

Boskin, M. 1973. Economics of the labor supply. In *Labor supply and income maintenance*, ed. G. Cain and H. Watts. Chicago: Rand McNally.

Brewer, M., A. Duncan, A. Shephard, and M.J. Suárez. 2006. Did working families' tax credit work? The impact of in-work support on labour supply in Great Britain. *Labour Economics* 13: 699–720.

Browning, M., A. Deaton, and M. Irish. 1985. A profitable approach to labor supply and commodity demands over the life-cycle. *Econometrica* 53: 503–544.

Burtless, G., and J.A. Hausman. 1978. The effect of taxation on labor supply: Evaluating the Gary negative income tax experiment. *Journal of Political Economy* 86: 1103–1130.

Cain, G. 1966. *Married women in the labor force*. Chicago: University of Chicago Press.

Card, D. 1994. Intertemporal labor supply: An assessment. In *Advances in econometrics, Sixth World Congress*, ed. C. Sims. New York: Cambridge University Press.

Card, D., and P.K. Robins. 1998. Do financial incentives encourage welfare recipients to work? Evidence from a randomized evaluation of the self-sufficiency project. In *Research in labor economics*, ed. S. Polachek, Vol. 17. Greenwich: JAI Press.

Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56: 63–90.

Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100: 437–467.

Chiappori, P.-A. 1997. Introducing household production in collective models of labor supply. *Journal of Political Economy* 105: 191–209.

Chone, P.-A., and G. Laroque. 2005. Optimal incentives for labor force participation. *Journal of Public Economics* 89: 395–425.

Cogan, J.F. 1980. Labor supply with costs of labor market entry. In *Female labor supply: Theory and estimation*, ed. J. Smith. Princeton: Princeton University Press.

Cogan, J.F. 1981. Fixed costs and labor supply. *Econometrica* 49: 945–964.

Donni, O. 2003. Collective household labor supply: Nonparticipation and income taxation. *Journal of Public Economics* 87: 1179–1198.

Eckstein, Z., and K. Wolpin. 1989. Dynamic labour force participation of married women and endogenous work experience. *Review of Economic Studies* 56: 375–390.

Eissa, N., and H. Hoynes. 2004. Taxes and the labor market participation of married couples: The earned income tax credit. *Journal of Public Economics* 88: 1931–1958.

Eissa, N., and J. Liebman. 1996. Labor supply response to the earned income tax credit. *Quarterly Journal of Economics* 111: 605–637.

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.

Fortin, B., and G. Lacroix. 1997. A test of the unitary and collective models of household labour supply. *Economic Journal* 107: 933–955.

Fraker, T., and R. Moffitt. 1988. The effect of food stamps on labor supply: A bivariate selection model. *Journal of Public Economics* 35: 25–56.

Gorman, W.M. 1959. Separable utility and aggregation. *Econometrica* 27: 469–481.

Gorman, W.M. 1968. The structure of utility functions. *Review of Economic Studies* 35: 367–390.

Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.

Gorman, W.M. 1961. On a class of preference fields. *Metroeconomica* 13, 53–56. Repr. in *Separability and Aggregation: Collected Works of M. Gorman*, vol. 1, ed. C. Blackorby and A.F. Shorrocks. Oxford: Clarendon Press, 1995.

Gronau, R. 1974. Wage comparisons: A selectivity bias. *Journal of Political Economy* 82: 1119–1144.

Gruber, J., and E. Saez. 2002. The elasticity of taxable income: Evidence and implications. *Journal of Public Economics* 84: 1–32.

Hall, R. 1973. Wages, income and hours of work in the US labor force. In *Income maintenance and labor supply*, ed. G. Cain and H. Watts. Chicago: Chicago University Press.

Ham, J. 1986. Testing whether unemployment represents intertemporal labour supply behaviour. *Review of Economic Studies* 53: 559–578.

Hanoch, G. 1965. The 'backward-bending' supply of labor. *Journal of Political Economy* 73: 636–642.

Hanoch, G. 1980. A multivariate model of labor supply. In *Female labor supply: Theory and estimation*, ed. J. Smith. Princeton, NJ: Princeton University Press.

Hanoch, G., and M. Honig. 1983. Retirement, wages, and labor supply of the elderly. *Journal of Labor Economics* 1: 131–151.

Hansen, L.P., and K. Singleton. 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91: 249–265.

Hausman, J. 1980. The effect of wages, taxes and fixed costs on women's labor force participation. *Journal of Pubic Economics* 14: 161–194.

Hausman, J. 1981. Labor supply. In *How taxes affect economic behavior?* ed. H. Aaron and J. Pechman. Washington, DC: Brookings Institution.

Hausman, J. 1985a. The econometrics of nonlinear budget sets. *Econometrica* 53: 1255–1282.

Hausman, J. 1985b. Taxes and labor supply. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, Vol. 1A. Amsterdam: North-Holland.

Hausman, J., and P. Ruud. 1984. Family labor supply with taxes. *American Economic Review* 74: 242–248.

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

Heckman, J.J. 1974b. Life cycle consumption and labor supply: An explanation of the relationship between income and consumption over the life cycle. *American Economic Review* 64: 188–194.

Heckman, J.J. 1974c. Effects of child-care programs on women's work effort. *Journal of Political Economy* 82: S136–S163.

Heckman, J.J. 1976. Life-cycle model of earnings, learning and consumption. *Journal of Political Economy* 84: S11–S44.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Heckman, J.J. 1993. What has been learned about labor supply in the past twenty years. *American Economic Review* 83: 116–121.

Heckman, J.J., and T.E. MaCurdy. 1980. A life-cycle model of female labour supply. *Review of Economic Studies* 47: 47–74.

Hotz, V.J., F.E. Kydland, and G.L. Sedlacek. 1988. Intertemporal preferences and labor supply. *Econometrica* 56: 335–360.

Hoynes, H. 1996. Welfare transfers in two-parent families: Labor supply and welfare participation under the AFDC-UP program. *Econometrica* 64: 295–332.

Hoynes, H. 2000. The employment, earnings and income of less skilled workers over the business cycle. In

Finding jobs: Work and welfare reform, ed. D.E. Card and R. Blank. New York: Russell Sage Foundation.
Hurd, M.D., and M.J. Boskin. 1984. The effect of social security on retirement in the early 1970s. *Quarterly Journal of Economics* 99: 767–790.
Immervol, H., H.J. Kleven, C.T. Kreiner, and E. Saez. 2007. Welfare reform in European countries: A microsimulation analysis. *Economic Journal* 117: 1–44.
Johnson, T.R., and J.H. Pencavel. 1984. Dynamic hours of work functions for husbands, wives and single females. *Econometrica* 52: 363–389.
Keane, M.P., and R. Moffitt. 1998. A structural model of multiple welfare program participation and labor supply. *International Economic Review* 39: 553–589.
Killingsworth, M.R. 1983. *Labor supply*. Cambridge: Cambridge University Press.
Killingsworth, M.R., and J.J. Heckman. 1986. Female labor supply: A survey. In *Handbook of labor economics*, ed. O. Ashenfelter and R. Layard, Vol. 1. Amsterdam: North-Holland.
Kooreman, P., and A. Kapteyn. 1990. On the empirical implementation of some game theoretic models of household labor supply. *Journal of Human Resources* 25: 584–598.
Krueger, A.B.., and J.-S. Pischke. 1992. The effect of social security on labor supply: A cohort analysis of the notch generation. *Journal of Labor Economics* 10: 412–437.
Laroque, G. 2004. Income maintenance and labour force participation. *Econometrica* 73: 341–376.
Lewis, H.G. 1967. *On income and substitution effects in labor force participation*. Unpublished manuscript, University of Chicago.
Low, H. 2005. Self-insurance and unemployment benefit in a life-cycle model of labour supply and savings. *Review of Economic Dynamics* 8: 945–975.
Lundberg, S. 1988. Labor supply of husbands and wives: A simultaneous equations approach. *Review of Economics and Statistics* 70: 224–235.
MaCurdy, T.E. 1981. An empirical model of labour supply in a life-cycle setting. *Journal of Political Economy* 89: 1059–1085.
MaCurdy, T.E. 1983. A simple scheme for estimating an intertemporal model of labor supply and consumption in the presence of taxes and uncertainty. *International Economic Review* 24: 265–289.
MaCurdy, T.E. 1985. Interpreting empirical models of labour supply in an intertemporal framework with uncertainty. In *Longitudinal analysis of labor market data*, ed. J.J. Heckman and B. Singer. Cambridge: Cambridge University Press.
MaCurdy, T.E. 1992. Work disincentive effects of taxes: A reexamination of some evidence. *American Economic Review* 82: 243–249.
MaCurdy, T.E., D. Green, and H. Paarsch. 1990. Assessing empirical approaches for analyzing taxes and labour supply. *Journal of Human Resources* 25: 415–490.
McElroy, M.B. 1981. Empirical results from estimates of joint labor supply functions of husbands and wives. In *Research in labor economics*, ed. R.G. Ehrenberg, Vol. 4. Greenwich: JAI Press.
Mincer, J. 1960. Labor supply, family income, and consumption. *American Economic Review* 50: 574–583.
Mincer, J. 1962. Labor force participation of married women: A study of labor supply. In *Aspects in labor economics*, ed. H.G. Lewis. Princeton: Princeton University Press.
Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.
Moffitt, R. 1983. An economic model of welfare stigma. *American Economic Review* 73: 1023–1035.
Moffitt, R. 1986. The econometrics of piecewise-linear budget constraints: A survey and exposition of the maximum likelihood method. *Journal of Business and Economic Statistics* 4: 317–327.
Moffitt, R. 2002a. Economic effects of means-tested transfers in the US. In *Tax policy and the economy*, ed. J.M. Poterba, Vol. 16. Cambridge, MA: MIT Press.
Moffitt, R.A. 2002b. Welfare programs and labor supply. In *Handbook of public economics*, ed. A.-J. Auerbach and M. Feldstein, Vol. 4. Amsterdam: North-Holland.
Moffitt, R. 2006. Welfare work requirements with paternalistic government preferences. *Economic Journal* 116: F441–F458.
Mroz, T.A. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55: 765–799.
Muellbauer, J. 1981. Linear aggregation in neoclassical labour supply. *Review of Economic Studies* 48: 21–36.
Pencavel, J. 1986. Labor supply of men: A survey. In *Handbook of labor economics*, ed. O. Ashenfelter and R. Layard, Vol. 1. Amsterdam: North-Holland.
Rust, J., and C. Phelan. 1997. How Social Security and Medicare affect retirement behavior in a world of incomplete markets. *Econometrica* 65: 781–831.
Saez, E. 2001. Using elasticities to derive optimal income tax rates. *Review of Economic Studies* 68: 205–239.
Saez, E. 2002. Optimal income transfer programs: Intensive versus extensive labor supply responses. *Quarterly Journal of Economics* 117: 1039–1073.
Scholz, J.K. 1996. In-work benefits in the United States: The Earned Income Tax Credit. *Economic Journal* 106: 156–169.
Shaw, K. 1989. Life-cycle labour supply with human capital accumulation. *International Economic Review* 30: 431–457.
Stern, N. 1986. On the specification of labor supply functions. In *Unemployment, search and labour supply*, ed. R.W. Blundell and I. Walker. Cambridge: Cambridge University Press.
van Soest, A. 1995. Structural models of family labor supply: A discrete choice approach. *Journal of Human Resources* 30: 63–88.
Wales, T.J., and A. Woodland. 1976. Estimation of household utility functions and labor supply response. *International Economic Review* 17: 397–410.

Labour Supply of Women

Mark R. Killingsworth

This article reviews theoretical and empirical work on the labour supply of women in modern times, with special reference to women in Western economies, primarily the United States.

The behaviour of female labour supply has important implications for many other phenomena, including marriage, fertility, divorce, the distribution of family earnings and male–female wage differentials. The labour supply of women is also of interest because of the technical questions it poses. For example, since many women do not work, corner solutions are potentially an important issue in both the theoretical and empirical analysis of female labour supply, even though in other contexts (e.g., studies of consumer demand) corner solutions are often ignored. (For recent discussions of this issue in the context of consumer demand studies, see Deaton 1986, and Wales and Woodland 1983.)

Female Labour Supply: Some Stylized Facts

Trends and Cyclical Patterns in Time-Series Data

Substantial secular increases in the labour force participation of women are a striking feature of the labour market in most developed economies in the 20th century. Growth in participation began at different times and has proceeded at different rates, but since the 1960s most advanced economies have seen considerable, and at times dramatic, rises in the proportion of women – particularly married women (especially those with small children) – in the labour force.

In both the US and Great Britain, participation rates of women have risen since 1890 for almost all individual age groups except those 65 or over, a pattern that, with a few exceptions, has been observed in most other Western countries as well (see Killingsworth and Heckman 1986, who provide extensive tabulations of many time series on female labour supply; and Sorrentino 1983). For example, between 1890 and 1980, the aggregate female participation rate in the US rose from 18.6 per cent to 50.5 per cent, and that for women 25–44 rose from 15.6 per cent to 64.9 per cent. Similarly, in Britain, the aggregate participation rate of women rose from 32.3 per cent in 1921 to 45.6 per cent in 1981, and that for women 25–44 rose from 28.4 per cent to 59.5 per cent during the same period.

Participation of married women is typically lower than that of single women. However, most of the recent *increase* in the aggregate female participation rate in the US, Britain, and other developed economies is attributable to an increase in the participation rate of married women. For example, between 1890 and 1980, the rate among married women in the US rose from 4.6 per cent to 50.1 per cent, whereas that for single women rose from 43.1 per cent to 61.5 per cent. Likewise, in Britain, the married female labour participation rate rose from 9.6 per cent to 47.2 per cent during 1911–1981, whereas the rate for single women actually declined somewhat, from 70.1 per cent to 60.8 per cent (with most of the decline occurring after 1961).

The substantial increase in participation among women, particularly married women, stands in sharp contrast with the secular decline in male participation rates. As Pencavel (1986) notes, male participation rates in developed economies have generally been falling – both in the aggregate and for most age groups – since at least the first quarter of the 20th century.

In contrast with the rise in participation rates, weekly hours worked by women workers in many Western countries (e.g., the US) appear to have been falling secularly. This decline in weekly hours worked by women workers parallels the decline in weekly hours worked by man that is documented by Pencavel (1986). For example, in 1940, about 40 per cent of employed women in the US worked more than 40 hours during the Census week, as opposed to only about 13 per cent in the 1980 Census.

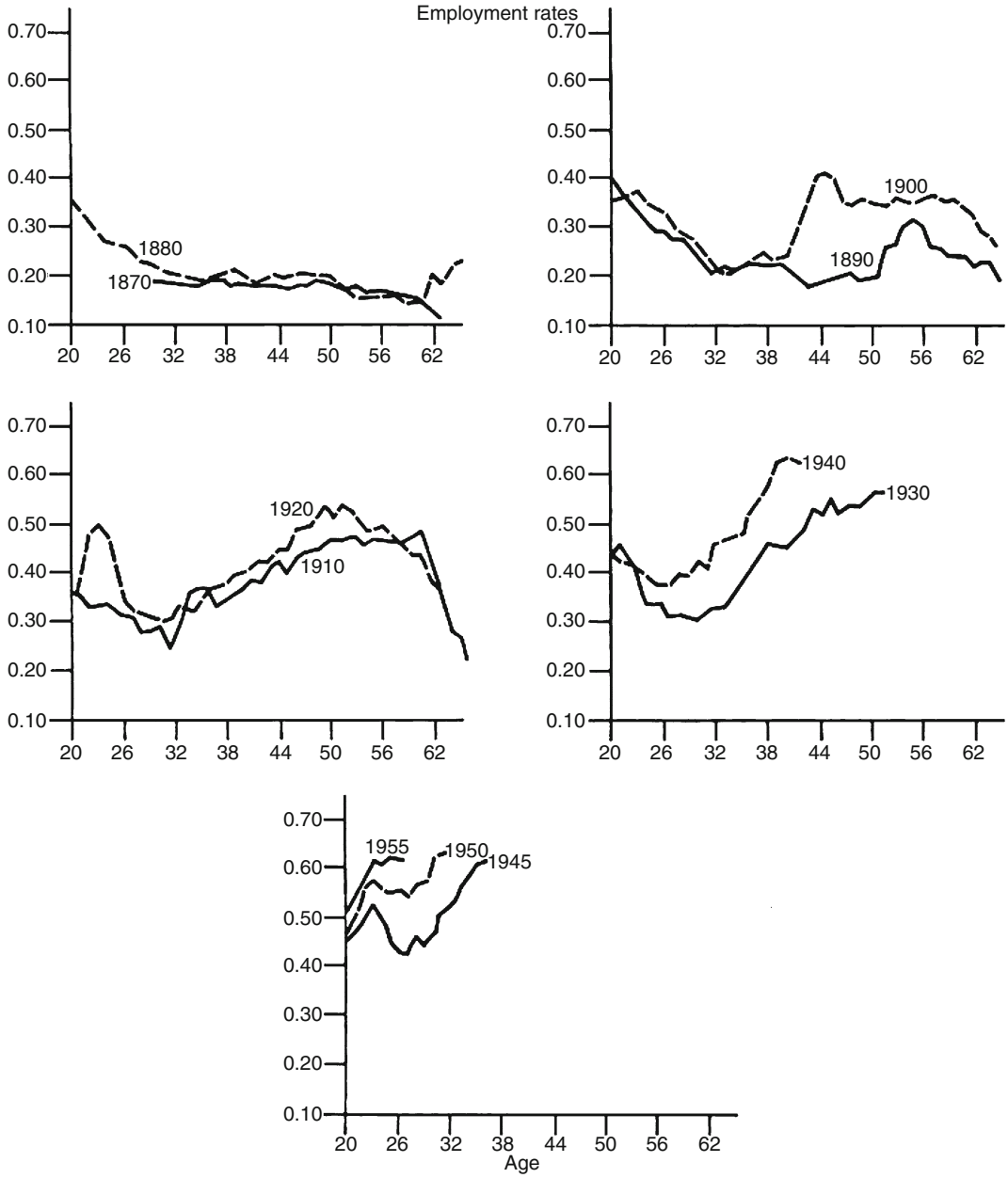
Considered alongside the substantial secular increase in women's participation rates, these secular reductions in hours of work raise several interesting questions. First, has the secular reduction in weekly hours worked by women workers been enough to offset the secular increase in the female participation rate and reduce the total number of hours of market work of women? One may address this question using Owen's (1986) constructed measure of 'total' weekly labour supply, 'labour input per capita', computed for US women as the product of the employment– population ratio and weekly hours worked by employed workers. Between 1920 and 1977, this measure of female labour input per capita approximately double among women age 25–64, increased slightly among women age 20–24 and declined only for the youngest (age 14–19) and oldest (65 or over) women.

Thus, the secular decline in female weekly hours worked has dampened, but has by no means fully offset, the effect of the secular increase in female participation in the labour force and in employment. On balance, the trend in total weekly labour input of women is clearly positive. Moreover, although participation and weekly hours of work are two of the most easily measured aspects of labour supply, they do not measure all aspects of labour supply. In particular, it is important to consider weeks worked per year as well.

The fact that weekly hours worked by women workers have fallen even as women's labour force participation has risen also poses a subtle question concerning within-cohort as opposed to across-cohort effects. The most obvious and straightforward interpretation of the secular decline in women's weekly hours of work is that hours worked per week by women workers have indeed fallen across successive cohorts. However, the decline in weekly hours worked has been accompanied by a substantial increase in participation, and this raises the question of whether the decline in weekly hours worked may be at least partly a consequence of the addition of 'lowhours' women, *within each cohort*, who would not be working had participation not increased. That is, if increased participation amounts to an influx of part-time workers (e.g., because greater availability of jobs with flexible hours has made work more attractive than before), then *average* hours worked may well fall even if hours worked by those *already* in the labour force stay the same or even rise.

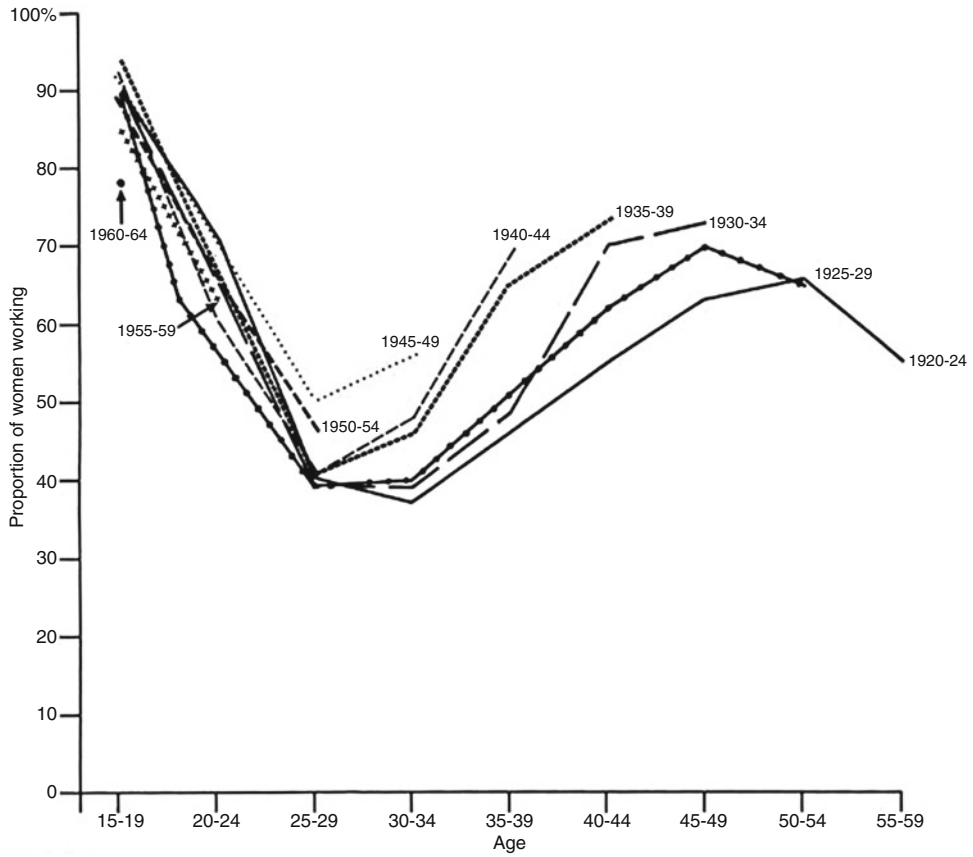
It is difficult to develop evidence on this issue: there are no data on the number of hours that a woman not now participating in the labour force *would* work *if* she were to work, much Fig. 1 Employment-Population Ratios by Age for Successive Female Birth Cohorts, 1870–1955, United States. Source: Smith and less data showing how this number has changed over time. It does, however, seem clear that successive cohorts of women have generally supplied steadily increasing amounts of labour, where 'labour supply' is defined as participation in the labour force, employment, weekly hours worked by the total population or annual hours worked (by either the working population or the total population). First, as shown in Fig. 1, participation in the labour force and in paid employment have increased in successive cohorts of US women: in general, more recent cohorts are more oriented towards market work than were earlier cohorts. Moreover, among the most recent cohorts there appears to have been a dampening or even a disappearance of the decline in market activity at childbearing and childrearing ages that was characteristic of earlier cohorts. Fig. 2 shows data on employment rates by cohort for Britain that tell a story similar to the one in Fig. 1, for the US.

Smith and Ward (1984, 1985) have derived two series on *annual* labour supply, by birth cohort, that provide additional evidence on these issues. (See also Smith 1983, who presents more detailed calculations for the shorter period 1977–81.) The first refers to annual hours worked by working women (calculated as the product of weekly hours worked times weeks worked per year among working women). It indicates that, at a minimum, *annual* hours worked by working women have not fallen at the same rate as *weekly* hours worked: evidently, the secular downtrend in the latter has been offset to a considerable extent by a secular increase in weeks worked per year.



Labour Supply of Women, Fig. 1 Employment –population ratios by age for successive female Birth Cohorts, 1870–1955, United States (Source: Smith and)

The second Smith–Ward cohort labour supply series provides analogous information on 'tota' annual labour supply, that is, the product of the employment–population ratio and annual hours worked by working women. Although the changes in total annual labour supply across cohorts are somewhat uneven, there is some indication that total annual labour supply is higher among more recent cohorts (though the growth in total annual labour supply, relative to earlier cohorts, is not nearly as dramatic as the increase in participation rates *per se*).



Labour Supply of Women, Fig. 2 Employment-population by age for successive female Birth Cohorts, 1920–60, Great Britain (Source: Martin and Roberts (1984, p. 119)).

The substantial increases in market work performed by women just noted have been the subject of considerable discussion and speculation for some time. In classic papers that helped inaugurate modern analysis of labour supply, Mincer (1962, 1963) suggested that secular increases in female labour supply could be interpreted, in part, as a substitution away from *nonmarket* labour ('housework'): unlike most men, who seemingly divide their time between leisure and market work and typically do relatively little housework, women could be regarded as having three main uses for their time – leisure, market work and nonmarket work. Subject to several technical caveats (Killingsworth and Heckman 1986, p. 135, n.7), it is then straightforward to apply the Hicks (1965, pp. 242–6) – Marshall (1920, pp. 386, 852–3) – Pigou (1946, p. 682) analysis of input demands to the demand for leisure: the elasticity of demand for a good (in this case, leisure) will be greater, the greater is the availability of alternatives to that good. Female leisure demand should be more elastic than male leisure demand because women have two alternatives to leisure whereas men have only one. Hence female leisure should respond more to wage changes than male leisure. Once one takes account of 'household technical progress' – labour-saving innovations in 'home production' such as washing machines, refrigerators, vacuum cleaners, frozen food and the like (Long 1958) – this argument provides a simple but seemingly compelling explanation of the dramatic secular increase in female labour supply.

A number of writers have challenged this view, however, particularly as regards nonmarket work. They argue that the amount of nonmarket work

performed by women has changed little, if at all, since the 1920s (see, e.g., Cowan 1983; Hartmann 1981; Vanek 1973, 1974), and suggest that the increase in female market work may have come primarily (if not entirely) at the expense of leisure.

Much of the evidence on this issue derives from the work of Vanek (1973, 1974), who considered time budget studies of fulltime housewives undertaken in the 1920s and 1930s and another time budget study for 1965, also of fulltime housewives, conducted by the University of Michigan Survey Research Center. According to Vanek, these studies suggest that housework performed by fulltime housewives has in fact changed remarkably little over 40 years.

Further reflection, however, raises questions about this conclusion. In view of the increase in female labour force participation rates, especially among married women, it seems clear that fulltime housewives are a declining (if by no means disappearing) species. Women who are (or remain) fulltime housewives despite a long-run trend towards participation of women, particularly married women, in market work may be an increasingly atypical segment of the female population. Comparisons over time with respect to this group may amount to a comparison of apples (the modal, or at least very frequent, behaviour pattern in the 1920s) and oranges (an increasingly less typical, albeit still important, behaviour pattern in the 1960s).

Moreover, careful examination of the available evidence, even if taken at face value, lends at most highly equivocal support to the claim that nonmarket work of fulltime housewives has changed little over time. The evidence for the period 1965–75 is in fact relatively clear on the opposite side of the issue: Survey Research Center studies indicate that female nonmarket work fell appreciably during these years (Owen 1986, p. 112; Robinson and Converse 1967; Robinson 1977).

There remains the evidence derived by Vanek (1973) for the period 1920–65. As Owen (1986, esp. p. 113) notes, much of this evidence is problematic. The time budget studies from the 1920s are unrepresentative in at least two important respects: most of them refer to farm women; and all of them appear to contain unrepresentatively large proportions of women in the higher social classes (e.g., women with high educational attainment and/or in families that owned their own farms). The upward class bias in the 1920s samples would understate the decline in female nonmarket work during 1920–65 to the extent that upper-class women in the 1920s did less nonmarket work than other women.

Moreover, even taken on its own terms, the evidence yields ambiguous conclusions about how female nonmarket work changed during 1920–65. 'Housework' – food preparation, clothing care, home care, etc. – in fact seems to have fallen substantially; other activities such as child care and shopping, which Owen (1986, p. 115) calls 'quasiwork', increased substantially. Narrowly defined so as to include only housework, nonmarket work seems to have fallen appreciably during 1920–65; only if one includes 'quasiwork' as part of nonmarket work is there any basis for the claim that nonmarket work was essentially unchanged over this period.

Although the quantitative changes in female labour supply noted above are quite remarkable, the 20th century has also seen striking qualitative changes in female labour supply, both in absolute terms and relative to men. In particular, in the US, the growth in the amount of female labour supply has been accompanied by a pronounced shift in its character: to a much greater extent than was true at the turn of the century, the representative woman worker today holds a white-collar – particularly a clerical – job. To some extent this simply reflects the economy-wide growth in the importance of white-collar work, but that is not the only factor, for the influx of women into white-collar (especially clerical) work occurred at a faster rate than did that of men.

For example, in 1900, 20.2 per cent of all women workers held white-collar (professional, technical, managerial, sales or clerical) jobs, vs. 65.6 per cent in 1980. Thus, the proportion of women in such jobs more than trebled over the period 1900–1980, whereas the proportion of men in such jobs increased by a factor of only about 2.4. The proportion of men in clerical jobs increased by a factor of about 2.3, whereas the

proportion of women in such jobs increased by almost ten-fold! Finally, the proportion of women in blue-collar (craft, operative, or labourer) and service jobs fell during 1900–1980 while the proportion of men in both kinds of jobs rose. Thus, both in absolute terms and relative to men, concentration of women in white collar (especially clerical) jobs rose whereas concentration of women in blue-collar and service jobs fell over the period 1900–1980.

I conclude this discussion of secular trends in female labour supply by briefly considering educational attainment, marital status and fertility. First, median educational attainment for successive cohorts of US women has increased only slightly in recent years (for example, in 1980, the median for the cohort born in 1926–30 was 12.3 years, vs. 12.8 years for the cohort born during 1951–5). However, over time, the education distribution has nevertheless changed considerably. For example, in 1980, 6.4 per cent of women born before 1906 had completed at least four years of college; the figures for the same year for women born during 1926–30 and 1951–5 are 9.9 per cent and 20.5 per cent, respectively.

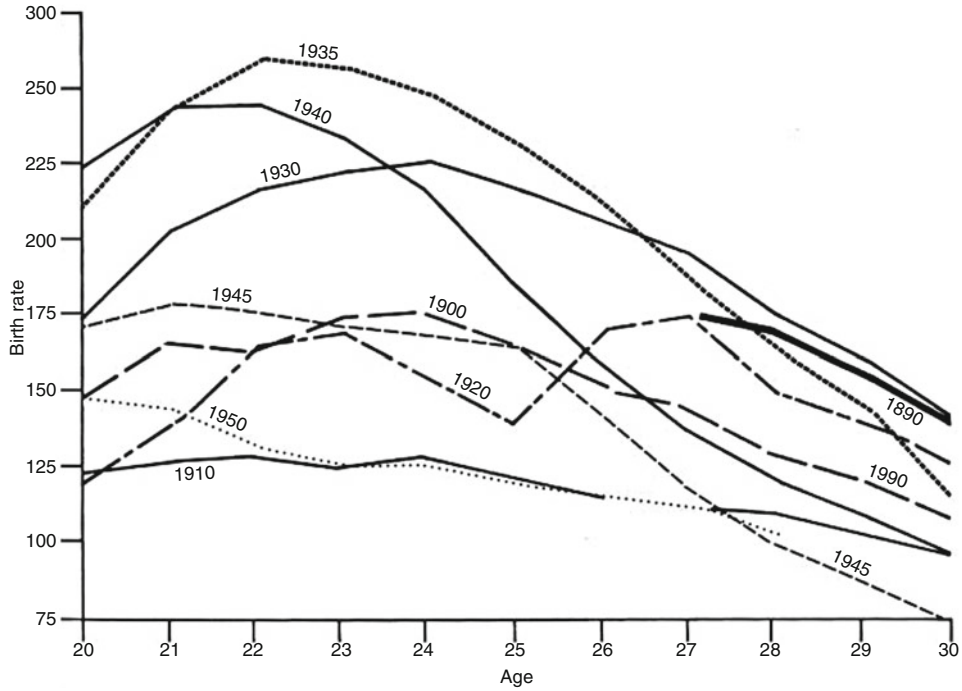
The distribution of women by marital status in the US has varied considerably during 1890–1980. The proportion of women in the 'other' category, consisting mainly of divorced women, has increased secularly (among women 25–29, 3.2 per cent were in this category in 1890, vs. 8.9 per cent in 1980), but otherwise the most noteworthy feature of women's marital status distributions in the US has been the degree to which they have fluctuated. For example, in 1980, the proportion never married and the proportion currently married among US women 25– 29 (20.8 per cent and 70.3 per cent, respectively) were both approximately equal to what they were in 1890 (25.4 per cent and 71.4 per cent, respectively), but each of these ratios has varied substantially during the period 1890–1980. For example, in 1960, 10.5 per cent of women then age 25–29 had never married and 83.3 per cent were currently married. Likewise, in both 1890 and 1980, slightly less than half of the women age 20–24 were married, but in 1960 almost 70 per cent of such women were married.

Figure 3 plots age-specific fertility rates for the ages between 20 and 30 for cohorts of US women between 1890 and 1950. As shown there, fertility rates rose substantially starting with the 1920 cohort (the 1910 cohort was age 20–30 during the Great Depression, which is probably a major reason why its fertility was below that of the 1900 cohort). However, starting with the 1940 cohort, fertility began to fall again; indeed, the pattern of fertility by age for the 1950 cohort was almost identical to that of the 1910 cohort.

Although I have often used the term 'trends' in discussing the time-series data on participation, schooling, etc., noted above, they in fact combine not only secular but also cyclical factors. For a rough and ready decomposition of observed time series into trend and cycle, one may regress first differences in the participation rate of a given female group (whites age 16–17, all nonwhites, etc.) on contemporaneous first differences in the unemployment rate of white males age 35–44, using annual data for 1955–1982. The intercept in these regressions (a) is an estimate of the secular trend in a given group's labour force participation rate, and the coefficient on the male unemployment variable (b) is a measure of the group participation rate's cyclical sensitivity.

The results of this exercise (Killingsworth and Heckman 1986, p. 122) are of some interest. In general, they indicate a strong secular uptrend in the participation rates of most female groups (as measured by the size and significance level of the intercept parameter, a), especially among whites. Most of the intercept or secular coefficients a are larger in absolute value than are the analogous coefficients for men in Pencavel (1986, Table 6). The results also suggest that female labour force participation is procyclical, in that the coefficient on the (change in the) male unemployment rate, b, is almost always negative and larger in absolute value than the coefficient derived by Pencavel for men in the same age group. However, in most cases this relation is imprecisely estimated for women and would not be called significant at conventional test levels.

Thus, these results and recent work by Clark and Summers (1981, 1982) and Coleman (1984) suggest that female labour force participation in the US is



Labour Supply of Women, Fig. 3 Age-specific birth rates for birth cohorts of 1890–1950, United States (Source: Smith and Ward (1984, p. 14))

not very sensitive to cyclical factors. (Joshi and Owen 1985, report similar findings for Britain.) In contrast, earlier work, most notably Mincer's (1966), found that participation – at least among married women – is strongly procyclical in the US. A major difference between Mincer's work and the more recent work is that the latter controls either implicitly or explicitly for possible serial correlation (e.g., by first-differencing, as in the work just discussed, or by maximum likelihood methods, as in Clark and Summers) whereas Mincer's work did not. Moreover, the recent results replicate Mincer's finding that the participation of teenage and prime-age women is relatively sensitive to cyclical variation; the finding of cyclical insensitivity in recent work has to do primarily with women age 45 or older.

Cross-section Patterns of Female Labour Supply

Most of the evidence on female labour supply discussed thus far refers to gross or unadjusted relationships between a measure of labour supply (e.g., labour force participation) and a single variable such as age or marital status. In this section, I briefly discuss relatively simple adjusted relationships between labour supply and such variables in cross-section, where 'adjusted' means that other factors have been held constant via simple statistical procedures. Although these adjusted relationships do not necessarily constitute a behavioural labour supply function, they do shed additional light on labour supply in the limited sense of documenting multivariate associations between labour supply and a number of variables of interest.

These relationships were derived by Bowen and Finegan (1969, esp. pp. 664–705), who consider labour force participation equations fitted to 1960 Census microdata for six different groups of single and married women in the age groups 25–54, 55–64 and 65–74 (the youngest group of married women includes women age 14–24 as well). Since Bowen and Finegan used least squares regression, their results may be interpreted as estimates of linear probability models.

In general, their results imply that labour force participation is strongly related to educational attainment, with greater schooling associated with increases (at a decreasing rate) in the probability of labour force participation. White single women below the age of 65 have a somewhat higher probability of participation than do black single women under 65, other things being equal; however, older white single women and all white married women have lower participation probabilities than do their black counterparts, ceteris paribus. Being (or having previously been) married is associated with a lower participation probability; so is having a large amount of 'other income' (i.e., income, including transfer income, other than own earnings), ceteris paribus.

The Bowen–Finegan results also suggest that, other things (including marital status and number of children) being equal, there is a fairly pronounced inverted-U-shaped relation between the probability of participation and age, especially among married women: Among younger women – single *or* married – being older is associated first with increased and then with reduced participation; among older women, participation tends to decline with age. Finally, for married women age 14–54 with spouse present, the presence of children (particularly children under the age of six) reduces the probability of participation.

Some Cautionary Remarks

Although this discussion has been concerned with stylized facts about labour supply, it should be noted, in conclusion, that the stylized facts presented here may not necessarily say much about structural, behavioural or 'casual' labour supply functions. Wage-hours combinations observed either in cross-section or over time do not necessarily trace out a behavioural ('causal') supply schedule. Rather, in general such data are the result of the interaction of both supply and demand. Thus, examination of stylized facts is only the beginning of a behavioural analysis, not the end. Accordingly, I now turn to theoretical models in labour supply and to empirical work aimed at deriving estimates of structural, behaviourally interpretable labour supply parameters.

Theoretical Models Pertinent to Female Labour Supply

Since the 1960s there has been an explosion of interest in labour supply generally and female labour supply in particular. In part, this interest was stimulated by the considerable changes in the labour force in the US and other countries described above; in part, it was encouraged by government funding of research on the labour-supply effects of transfer programmes (notably the so-called negative income tax experiments: see, e.g., Moffitt and Kehrer 1981, 1983). One important result of the research conducted since the early 1960's has been an array of new theoretical labour supply models. Full details of these models have been summarized elsewhere (see, e.g., Heckman et al. 1981; Killingsworth 1983; Killingsworth and Heckman 1986; Pencavel 1986); in what follows, I limit my discussion to labour supply models that are or might be especially pertinent to analysis of female labour supply and to understanding the patterns described earlier.

In broad terms, even the simplest labour supply model of Robbins (1930) and Hicks (1946) can account for some of the most important stylized facts about the behaviour of female labour supply. In that model, two key economic variables affect labour supply: the real wage, and real 'exogenous' income (i.e., income from sources unrelated to one's own work, such as income from property). Empirically, labour supply responses to changes in exogenous income appear to be small in relation to responses to changes in real wages (which have usually – though by no means always – been found to be positive among most women). Hence, secular growth in the real wage of women would be expected to increase female labour supply notwithstanding secular growth in 'exogenous' income (which might be interpreted as including earnings of husbands as well as income from property and the like). (See Mincer 1985, for an attempt to use crosssection labour supply parameter estimates to explain the time-series behaviour of female labour supply.)

More elaborate models of labour supply have the potential to provide a richer understanding of

patterns and secular trends in female work effort – although, as noted below, such models often raise more questions about female labour supply than they answer. The previous section suggests that several phenomena – e.g., marriage and family membership, nonmarket work, the occupational dimension of labour supply, childbearing and childrearing (and the intertemporal planning issues that they raise) – are of special interest in discussions of female market work. In this section, I consider models pertinent to each of these matters.

Marriage, Family Membership and Nonmarket Work

Marriage and family membership, and the obligations that accompany them, seem to be very important correlates of levels of and trends in female labour supply. (For example, for married women the *level* of labour supply is generally lower but the positive *trend* has generally been higher than for single or other women.) It would therefore seem that models that explicitly recognize important economic aspects of marriage and family membership would enhance understanding of female labour supply.

In the conventional family labour supply model, a single decisionmaking unit, the family or household, maximizes a *family* or *household* utility function (whose arguments are total family consumption and the leisure times of each of the family's members) subject to the constraint that total family or household income (exogenous income plus all family member' earnings) may not exceed total family expenditure on goods. This model may be regarded as an extension of the simple Robbins–Hicks labour supply analysis (which may best be thought of as a treatment of the labour supply decision of a single individual) or, alternatively, as a version of the standard model of the consumer's choice of n distinct consumption goods (with the decision about purchases of Apples, Bread,. . .;, converted into a decision about the leisure consumption and labour supply of Alfred, Bertha,. . .;). Thus the standard results of consumer theory carry over to the family labour supply model with little or no essential modification. (For econometric work based on the family labour supply model, see Ashenfelter and Heckman 1974, and Hausman and Ruud (1984).

Since (by assumption) the family has a common utility function and pools its income, a change in one family member's wage or exogenous income will affect not only his or her own behaviour but also that of other family members. Such intrafamily substitution of labour and leisure is obviously one of the most important implications of the conventional family labour supply model. Of particular interest here are the intrafamily adjustments that may occur when some (but not all) family members are 'rationed' (i.e., constrained from offering all the market work, or consuming all the leisure, that they would otherwise choose to do). Such rationing may entail one of the members being unemployed or, alternatively, at a corner solution (devoting all available time to leisure) and may have various consequences. One is the so-called 'added worker effect', whereunder unemployment of the husband tends to increase the probability that his wife will enter the job market (see Ashenfelter 1980; Lundberg 1985; and Mincer 1966, for further discussion). Another set of implications of such rationing concerns the difference in behaviour between households with working wives and those with nonworking wives. For example, given suitable assumptions, one can show that (i) the male compensated substitution effect will be smaller in families with nonworking wives, (ii) the income effect on household consumption will be larger (smaller) for households with working wives if the wife's home time and consumption goods are net substitutes (complements) and (iii) the compensated or crosssubstitution effect of a rise in the male wage on demand for consumption goods will be smaller (larger) in families with nonworking wives if one spouse's leisure is a net substitute for market goods whereas the other's is a not complement (if the spouses' leisure times are *both* either net complements *or* net substitutes with market goods). (See Heckman 1971, Essay III; Killingsworth and Heckman 1986, p. 130; and Kniesner 1976.).

As most married people will readily testify, nonmarket work is an important aspect of family

life; and much of it is done by women. The previous section's discussion of trends in women's nonmarket work suggests that models of family decisionmaking that explicitly account for nonmarket work may provide an explanation both for the fact that the *level* of market work is lower for women than for men and for the frequent (but – see below – by no means universal) empirical finding that the *elasticity* of market work with respect to wage rates is greater among women than among men.

Most analyses of the relations between family market work, nonmarket work and leisure derive from the time allocation model of Becker (1965). In this approach, the basic objects of choice are not consumer goods and leisure times, but rather 'commodities' or 'activities' that a 'produced' using 'inputs' of market goods and family members' times subject to 'household production functions': for example, cooking utensils, raw food and the time of one or more family members produce a cooked meal. (In this connection, it is instructive to note that Leibowitz 1974, pp. 246–7, reports that husbands' and wives' times are substitutable in meal production at the marginal rate of ten minutes of husband time for each five minutes of wife time!)

If wives have a comparative advantage at *nonmarket* production (i.e., a higher elasticity of output with respect to time input) relative to husbands, then it can be shown that, in general, wives will tend to specialize at nonmarket production even if they can earn the same wage doing market work as husbands. Hence the *level* of labour supply will be lower for wives than for husbands; and the existence of more alternatives to leisure (i.e., nonmarket as well as market work) will tend to make the *elasticity* of labour supply with respect to wage rates greater for wives than for husbands. These conclusions are reinforced if the wife's market wage is less than the husband's. (See Graham and Green 1984; and Killingsworth and Heckman 1986, p. 138.)

At least in these respects, then, the time-allocation version of the family labour supply model seems to provide a strikingly successful account of female work effort. However, this success may be more apparent than real. First, nothing in the model requires that the greater allocation of wife's (as opposed to husband's) time to nonmarket work be a result of comparative advantage in the technical sense; just the same results would arise if the household were simply biased towards using the wife's time in nonmarket work for reasons (psychological, cultural, etc.) that have nothing to do with technological production possibilities per se. Indeed, one could get the same results by ignoring time allocation considerations entirely and by instead simply assuming a conventional household utility function that is biased towards female leisure time (Killingsworth and Heckman 1986, pp. 138–9). Perhaps most important, although the household behaviour model posits a household utility function (without specifying where it comes from), families may grow or dissolve, and in any case are made up of individuals. Thus the case for a household utility function has proven to be somewhat difficult to argue on *a priori* grounds. Perhaps in part for this reason, some recent work has sought alternatives to the household utility function approach. In some cases (e.g., Ashworth and Ulph 1981; Bourguignon 1984; Kooreman and Kapteyn 1985; Leuthold 1968) family members are assumed to pool their incomes for purposes of consumption and to maximize their *individual* utility (which depends on their own leisure time and on *family* consumption, which is thus effectively taken to be a public good) subject to a constraint on total *family* expenditure. This approach is formally very similar to the analysis of product-market duopolists who maximize their own profits but share the same market (Allen 1938, esp. pp. 200–204).

In other cases (notably Horney and McElroy 1978; Manser and Brown 1979, 1980; and McElroy and Horney 1981), decisions of individual family members – and, for that matter, the existence of the family itself – are treated in game-theoretic terms. For example, McElroy and Horney (1981) develop a Nash-bargained system of labour supply and commodity demand equations for each individual in a two-person family as the result of a constrained static, nonzero-sum game. Bargaining models of this kind have several interesting features. Since they emphasize

individuals' decisions and explicitly allow for alternatives to marriage, such models can be used in analyses of marriage and divorce. Also, since such models emphasize the bargaining power of individual family members, each individual family member's exogenous income appears as a separate argument in each demand equation (for leisure, consumption, etc.), and – shades of certain Victorian novelists! – changes in the intrafamily distribution of wealth will affect family members' bargaining strengths and, thus, their behaviour.

Unfortunately, empirical work on these and similar rivals to the conventional household utility function approach is still in its early stages. One problem inherent in such work is that variables that play a key role in bargaining – e.g., exogenous income flows that are under the control of each specific family member – are generally not measured in available data.

In principle, alternatives to conventional economic paradigms might yield results of interest in both empirical and theoretical analyses of family membership and its implications for female labour supply (and, more generally, the economic role of women). In practice, however, this does not yet appear to have happened. The rather small quantum of research on the subject undertaken by Marxists and other non- or anti-neoclassicals (e.g., Beneria 1977; Himmelweit and Mohun 1977; Humphries 1977) does not seem to have produced new insights, since it has been concerned more with description (e.g., of household production in Marxist terms) than with generation of testable hypotheses. Similarly, the best-known radical feminist work in this area (Hartmann 1981) discusses descriptive statistics on the sizeable female–male differential in housework time and emphasizes the family as a locus of 'struggle', but overlooks more sophisticated empirical work (e.g., Gronau 1973a, b, c, 1977; Leibowitz 1974) and the bargaining models noted above.

Costs of labour market entry. Popular discussions of women's work (particularly work by women with small children) often emphasize the important role of 'costs of labour market entry' – especially for child care and related services. A man (person?) from Mars might have difficulty understanding why households think of child care as a cost of the *wife's* (as opposed to the husband's) entering the labour market. However, casual observation suggests that at least some households do in fact think of child care and related costs in this way; and any unmarried person (male or female) with dependent children will obviously have to consider the cost of child care in deciding on whether to work in the paid labour force (even if the *ex post* level of such cost is zero, i.e., even if the children are left to fend for themselves).

To consider some of the implications of these costs, ignore family complications and focus solely on individuals (e.g., an unmarried mother with one or more dependent children); and treat the level of such costs as exogenously given (although at least some of these costs may well be the result of an optimizing decision subject to constraints). If child care services can be purchased (or other labour market entry costs incurred) on a strictly per-hour basis (as with, e.g., baby sitters and child-minders), then such costs are the equivalent of a reduction in the individual's hourly wage. They can therefore be expected to reduce the probability that a given individual will participate in the paid labour market and will have the usual positive income and negative substitution effects on hours worked by those who do in fact work for pay. A more interesting case arises when such costs are at least to some extent fixed or 'lumpy' (e.g., as when the individual must pay a fixed sum for a fixed number of hours of child care, as with nursery schools). Such costs induce a nonconcavity in the individual's budget set (if the amount of such costs is C whereas nonwork income is N, then the individual's income if she does no work is N but her income after just the first minute on the job is N – C (plus a minute's wages)); moreover, since by definition they are not affected (within the relevant range) by the amount of work the individual does, such costs are the equivalent of a reduction in network income (rather than a reduction in the hourly wage). Provided leisure is a normal good, an individual will be less likely to participate in the paid labour market, the higher

are such costs; *but* any working individual will work more hours, the higher are such costs. (See Heckman 1974, and Killingsworth 1983, esp. pp. 23–8, for elaboration, including discussion of alternative kinds of subsidies for child care and other costs of labour market entry.)

Labour Supply with Heterogeneous Jobs

Although changes in the amount of work done by women have been accompanied by substantial changes in the *kind* of work done by women (recall the discussion of the changing occupational composition of the female work force in section "Female Labour Supply: Some Stylized Facts," above), surprisingly little has been done to allow for heterogeneous types of jobs in the analysis of labour supply. At least two approaches are possible. The first, developed by Atrostic (1982), considers a finite number of *job characteristics* possessed in varying degrees by each of a potentially infinite number of jobs. Utility depends on consumption, leisure and the amount of each characteristic at one's current job; as in the literature on compensating wage differentials, the wage rate (and thus the budget constraint) is likewise a function of job characteristics. This approach leads conveniently to a model that closely resembles those used in estimating consumer demands: in effects, each job characteristic is treated as an endogenously chosen 'good'.

The second approach to analysing labour supply to heterogeneous jobs considers the discrete choice among a finite number of jobs each of which possesses varying degrees of a potentially infinite number of characteristics (Hill 1985; Killingsworth (1985). Not only the wage rate but also the utility function (or indirect utility function, etc.) is job-specific: the wage that one can earn with a given set of characteristics (educational attainment, prior work experience, etc.) will be different in different jobs; and the utility that one can derive from a given bundle of consumption and leisure will depend on the job one is doing – i.e., on where one spends one's *non*leisure time. One chooses the particular job that yields the highest possible value of utility; labour supply to that job is then determined in the usual way (e.g., by direct maximization of the utility function specific to the job in question subject to the budget constraint, with the job-specific wage; or by direct application of Roy's Identity to the jobspecific indirect utility function).

Unfortunately, empirical work on such models is even scarcer than empirical work on family bargaining models. Although their methodology is clearly relevant to female labour supply, the studies by Atrostic (1982) and Killingsworth (1985) are concerned with male work effort. Hill (1985), however, uses a discrete job choice model to analyse the decision of Japanese women to work in the informal family firm sector, the formal ('employee') sector or to work exclusively in the home. Application of such models to female labour supply in other settings is a potentially important area for future research.

Dynamic Issues

Although the discussion thus far has focused on models of an essentially static nature, much recent work has emphasized that the labour supply decision generally, and the labour supply of women in particular, raises questions of a fundamentally dynamic nature. The most noteworthy example concerns analyses of women's wages and of sex differentials in wages, in which the life cycle pattern of labour supply – e.g., the role of intermittent or continuous participation in the job market – and human capital investment decisions have often been assigned a crucial role. A long tradition in discussions of women's behaviour over the life cycle (exemplified by, e.g., Mincer and Polachek 1974), which I will call the 'Informal Theory', identifies the age of childbearing and childrearing as a period of reduced *investment in human capital* as well as of reduced *labour supply*; and links the low level of (or rate of growth in) women's wages during this period to the hypothesized low level of such human capital investment. More elaborate versions of the Informal Theory emphasize the long-run considerations underlying investment and labour supply decisions, stressing, for example, that women who anticipate a period of absence from the labour force in the future (for, e.g., childbirth and childrearing) may invest relatively little in skill

enhancement in the present (see, e.g., Mincer and Ofek 1982; Mincer and Polachek 1978; Polachek 1979, 1981).

Although the Informal Theory has considerable intuitive appeal, its very informality raises an important issue. Simple reasoning based on the Informal Theory often proceeds as if causation ran from (low) labour supply to (low) investment in skills; but in a long run perspective both investment and labour supply are choice variables, determined by other, more fundamental forces. What are these forces, and how may they be modelled? Here the Informal Theory is not particularly specific, or, therefore, especially helpful. In an attempt to spell out more clearly what the Informal Theory is (or could be) saying, Killingsworth and Heckman (1986) extend a conventional model of life cycle labour supply and human capital accumulation (Heckman 1976) by introducing a 'taste shifter' variable $m(t)$ into the utility function. This taste shifter affects the amount of 'effective' leisure in the utility function in the same way that technical progress affects an input in the production function: a high (rising) $m(t)$ denotes a large (or growing) taste for leisure time at time t. Thus introduction of $m(t)$ is a simple way to represent explicitly (if crudely) the notion that, for a variety of reasons (cultural, biological, etc.), a given woman's desire for 'leisure' (for, e.g., childbearing and childrearing) may change over time; and the related notion that, at any given date, different women will for various reasons have different preferences for such leisure.

The implications of the analysis may be discussed under two heads: equilibrium dynamics and comparative dynamics. The former refers to 'evolutionary' changes in labour supply, wages, etc., over time as a women follows an intertemporal equilibrium path in fulfillment of a lifetime plan formulated with respect to specific values of the 'givens' of the mode (including m) at each date; the latter, to changes (or cross-section differences) in labour supply, wages, etc., *at given dates* in response to 'parametric' *changes* (or differences) in the underlying givens of the model – e.g., initial asset holdings, the value of m at a given date, etc.

As regards equilibrium dynamics, to the extent that $m(t)$ can legitimately interpreted as an index of women's greater preference for nonmarket time during the age of childbearing and childrearing, the model provides a comprehensive and seemingly quite appealing set of predictions about life-cycle patterns of work and wages for women. It implies that, ceteris paribus, leisure will be higher (or rising more rapidly), and both the wage rate and labour supply will be lower (or rising less rapidly), during the childbearing and childrearing ages than at other points in the life cycle. In broad terms, these are clearly consistent with the stylized facts about the age pattern of female labour supply noted in section "Female Labour Supply: Some Stylized Facts."

In other respects, however, the model's equilibrium dynamics implications seem at odds with the Informal Theory. In particular, high or growing $m(t)$ is predicted *not* to affect investment time or the human capital stock at all; and the 'investment content' of time spent at work – i.e., the extent to which an hour of work time contributes to the accumulation of skills – is predicted to be relatively high during the age of childbearing or childrearing even though the total *amount* of time spent at work is predicted to be low.

Similar puzzles emerge from the model's comparative dynamics properties (which are in effect propositions about cross-section differences between women with different characteristics, e.g., different values of m at a given date t). Here it appears that, *during the childbearing and childrearing ages*, women with a greater taste for nonmarket work will tend to have (i) *lower* hours of work and wage rates, and (ii) *higher* hours of leisure and a higher investment content per hour spent at work, ceteris paribus. However, the model also suggests that, at ages *other than* those of childbearing and childrearing, these patterns will be precisely reversed: then, women with a greater taste for nonmarket work during the childrearing years (i.e., those who anticipate subsequent reduction of market work) will spend *more* time working, receive higher earnings per hour spent at work, devote less time to leisure and work at jobs with a relatively low investment content. Thus, although formal analysis and the

Informal Theory are basically in agreement on some of the main questions about behaviour during the childbearing and childrearing years, they disagree on others (e.g., the 'investment content' of work during those years); and the formal analysis highlights something ignored by the Informal Theory, i.e., an implicit substitution between periods with high and low *m*.

Empirical Analyses of Female Labour Supply

I now discuss empirical analyses of female labour supply, focusing on work based on static models (estimation using dynamic models is still in its infancy). To motivate this discussion, it is worth noting at the outset that the results of some recent work differ appreciably from those of research undertaken through the early 1980s. There has been a consensus of relatively long standing that compensated and uncompensated female labour supply wage elasticities are positive and larger in absolute value than those of men. In contrast, some recent studies suggest that elasticities for women differ little from those of men; indeed, in this work, the female uncompensated wage-elasticity of labour supply is often estimated to be negative.

Methodological Issues

Some of the most interesting aspects of empirical work on female labour supply have to do not with substantive findings but, rather, methodological innovations. 'First-generation' research on female labour supply, which proceeded through roughly the mid-1970s, approached empirical analysis of female labour supply using a conventional least squares regression framework: hours of work were regressed on variables denoting the wage rate, exogenous income and a vector of other (e.g., demographic) characteristics. The difficulty with this is that it ignores the implications of virtually all theoretical labour supply models. In particular, in such models, wages, exogenous income and other factors will of course have *no* effect on labour supply unless labour supply is positive – or, equivalently, unless the wage rate exceeds the 'reservation wage' (the lowest wage at which an individual would be willing to work). Otherwise, the derivative of labour supply with respect to any variable – the wage, exogenous income, demographic characteristics – is zero. The conventional regression approach ignores this fundamental notion and thus misspecifies the labour supply function. Development of a more comprehensive approach, one that accounts both for the decision to work and for the hours worked by persons who are working, has been a central feature of subsequent 'second-generation' research (see, e.g., Killingsworth 1983).

Empirical work on female labour supply has also had to confront the fact that, since many women are not working at any given moment, data on the market wages of nonworking women (i.e., the wage such women would be capable of earning if they were to work) are not available. Thus, analysis of the decision to work is more difficult than would otherwise be the case. It might seem (and, to many first-generation researchers, did in fact seem) that the simplest way to avoid both these problems is to fit labour supply functions to data restricted to *working* women: Among working women, changes in ways and other variables will generally have non-zero effects on labour supply; and data on wages are generally available for such women. However, this solution raises a new problem of an econometric nature, variously called 'sample selection' or 'selectivity' bias: If working women are not representative of *all* women, then least squares regression analysis of data restricted to working women may induce bias in estimates of structural (e.g., utility function) parameters. A simple intuitive argument suggests the nature of the problem. More or less by definition, working women are women for whom the wage (the 'offered' or 'market' wage), *w*, exceeds the 'reservation' level *w**. Thus, among all women who can earn a given market wage rate *w*, working women have relatively low *reservation* wages *w**; and, by the same token, among all women with a given reservation wage, *w**, working women must have relatively high *market* wages, *w*. On both counts, then, working women are unlikely to be unrepresentative of the entire female population.

Since the objective of empirical analysis is usually to derive estimates of population parameters (e.g., the parameters of utility functions), this bodes ill for conventional least squares regression, in which the error term is assumed to be a mean-zero variable uncorrelated with the regressors (e.g., the wage rate and exogenous income). Development of alternative econometric strategies to cope with this issue has been an important concern of much second-generation work on female labour supply (Heckman and MaCurdy 1986; Killingsworth 1983, ch. 3; Wales and Woodland 1980).

Empirical Findings

Although second-generation research has done much to enhance the intellectual rigour of empirical analysis of female work effort, it has not produced a consensus on the magnitudes of female labour supply parameters. Many of the estimates of the gross ('uncompensated') elasticity of female labour supply with respect to the wage rate are in the range 0.5–1.0 (see the summary in Killingsworth and Heckman 1986), which is rather large in absolute terms and very large relative to male elasticities (see, e.g., the summary in Pencavel 1986). However, the variance in these estimates is substantial. For example, Dooley (1982) and Heckman (1980) have obtained elasticity estimates in excess of +14.00 (!), whereas for other groups of women Dooley (1982), Nakamura and Nakamura (1981), and Nakamura, Nakamura and Cullen (1979) have derived estimates of −0.30 or less.

To some extent, the diversity of female labour supply parameter estimates is a direct consequence of the diversity of newly-developed datasets, theoretical models and econometric techniques. Sensitivity analyses that highlight the marginal effects of adopting different specifications or econometric procedures for the same dataset therefore seem necessary, or at least potentially very useful, for sorting out some of the reasons for the substantial variation in estimates. The most elaborate sensitivity study now available, that of Mroz (forthcoming), offers some surprising results that challenge the received view that female labour supply elasticities are generally rather large but does not, unfortunately, resolve all questions about the different results obtained in different studies.

For example, Mroz uses 1976 Panel Study of Income Dynamics (PSID) data on white wives age 30–60 to replicate – with the same variables and statistical procedures – work by Heckman (1980), who analysed data on white wives age 30–44 in the 1966 National Longitudinal Survey (NLS). Mroz's estimates of the uncompensated wage-elasticity of female labour supply are uniformly lower than Heckman's. Adding variables not included in Heckman's analysis results in greater elasticity estimates, but it also raises the estimates' standard errors. Possibly, the Mroz and Heckman estimates differ because they come from different datasets (the 1966 NLS vs. the 1976 PSID): although mean hours worked by *working* women are about 1300 hours per year in both datasets, participation rates are quite different (0.36 for Mroz, 0.47 for Heckman). However, the difference in datasets is probably not the whole story. For example, Cogan (1980), like Mroz, uses the 1976 PSID (albeit for essentially all white wives regardless of age, as opposed to Mroz's smaller group of white wives age 30–60) and gets an implied wageelasticity of 1.14, much higher than most of Mroz's estimates.

In sum, although recent work has provided a firmer methodological base for empirical analyses of female labour supply, it has raised more questions than it has answered about the actual magnitudes of the parameters governing work effort of women. There is a silver lining to this cloud, however: researchers interested in female market work are unlikely to run out of things to do for the foreseeable future.

Note on the Literature

The literature on female labour supply is substantial. Heckman (1978) and Killingsworth and Heckman (1986) discuss theoretical models and empirical results; the text by Blau and Ferber (1986) presents much useful material. Smith, ed. (1979), gives a general overview of women in the US labour market; Fuchs (1984),

Goldin (1980, 1983a, b, 1984, 1986), Goldin and Sokoloff (1982) and Smith and Ward (1984, 1985) discuss historical and recent trends. The papers in Layard and Mincer (eds, 1985), include work on female labour supply in Australia, Britain, the Federal Republic of Germany, France, Israel, Italy, Japan, the Netherlands, the Soviet Union, Spain, Sweden and the US. See also Joshi (1985), Joshi and Owen (1984, 1985), and Martin and Roberts (1984) on Britain; Nakamura and Nakamura (1981), Nakamura, Nakamura and Cullen (1979), Smith and Stelcner (1985), Stelcner and Breslaw (1985), Stelcner and Smith (1985) and Robinson and Tomes (1985) on Canada; Franz (1981) and Franz and Kawasaki (1981) on the Federal Republic of Germany; Bourguignon (1985) on France; Hill (1983, 1984, 1985), Yamada and Yamada (1984, 1985) and Yamada, Yamada and Chaloupka (1985) on Japan; and Kapteyn, Kooreman and van Soest (1985), Kooreman and Kapteyn (1984, 1985), Renaud and Siegers (1984) and van der Veen and Evers (1984) on the Netherlands.

See Also

- ▶ Discrete Choice Models
- ▶ Family
- ▶ Gender
- ▶ Household Production
- ▶ Segmented Labour Markets
- ▶ Selection Bias and Self-Selection
- ▶ Value of Time
- ▶ Women and Work
- ▶ Women's Wages

Bibliography

Allen, R.G.D. 1938. *Mathematical analysis for economists*. London: Macmillan.

Ashenfelter, O. 1980. Unemployment as disequilibrium in a model of aggregate labor supply. *Econometrica* 48: 547–564.

Ashenfelter, O., and J.J. Heckman. 1974. The estimation of income and substitution effects in a model of family labor supply. *Econometrica* 42: 73–85.

Ashworth, J.S., and D.T. Ulph. 1981. Household models. In *Taxation and labour supply*, ed. C.V. Brown, 117–133. London: Allen & Unwin.

Atrostic, B.K. 1982. The demand for leisure and nonpecuniary job characteristics. *American Economic Review* 72: 428–440.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Beneria, L. 1977. Reproduction, production and the sexual division of labour. *Cambridge Journal of Economics* 1: 203–225.

Blau, F.D., and M.A. Ferber. 1986. *The economics of women, men, and work*. Englewood Cliffs: Prentice-Hall.

Bourguignon, F. 1984. Rationalité individuelle ou rationalité stratégique: le cas de l'office familiale de travail. *Revue Economique* 35: 147–162.

Bourguignon, F. 1985. Women's participation and taxation in France. In *Unemployment, job search and labour supply*, ed. R. Blundell and I. Walker, 243–266. Cambridge: Cambridge University Press.

Bowen, W., and T.A. Finegan. 1969. *The economics of labor force participation*. Princeton: Princeton University Press.

Clark, K.B., and L.H. Summers. 1981. Demographic differences in cyclical employment variation. *Journal of Human Resources* 16: 61–79.

Clark, K.B., and L.H. Summers. 1982. Labour force participation: Timing and persistence. *Review of Economic Studies* 49(Supplement): 825–844.

Cogan, J. 1980. Labour supply with costs of labor market entry. In *Female labor supply*, ed. J.P. Smith, 327–364. Princeton: Princeton University Press.

Cowan, R.S. 1983. *More work for mother*. New York: Basic Books.

Deaton, A. 1986. Demand analysis. In *Handbook of econometrics*, vol. 3, ed. Z. Griliches and M. Intriligator. New York: North-Holland.

Dooley, M.D. 1982. Labor supply and fertility of married women: An analysis with grouped and individual data from the 1970 U.S. Census. *Journal of Human Resources* 17: 499–532.

Franz, W. 1981. Schatzung regionaler Arbeitsangebotsfunktionen mit Hilfe der Tobit-Methode und des Probit-verfahrens unter Berucksichtigung des sog. 'Sample Selection Bias'. Discussion Paper No. 171–81, Institut für Volkswirtschaftslehre und Statistik, University of Mannheim, Mannheim, Federal Republic of Germany.

Franz, W., and S. Kawasaki. 1981. Labor supply of married women in the Federal Republic of Germany: Theory and empirical results from a new estimation procedure. *Empirical Economics* 6: 129–143.

Fuchs, V. 1984. His and hers: Gender differences in work and income, 1959–1979. Working Paper No. 1501, National Bureau of Economic Research, Cambridge, MA.

Goldin, C. 1980. The work and wages of single women, 1970 to 1920. *Journal of Economic History* 40: 81–88.

Goldin, C. 1983a. The changing economic role of women: A quantitative approach. *Journal of Interdisciplinary History* 13(4): 707–733.

Goldin, C. 1983b. Life cycle labor force participation of married women: Historical evidence and implications.

Working Paper No. 1251, National Bureau of Economic Research, Cambridge, MA.
Goldin, C. 1984. The historical evolution of female earnings functions and occupations. *Explorations in Economic History* 21: 1–27.
Goldin, C. 1986. Monitoring costs and occupational segregation by sex: A historical analysis. *Journal of Labor Economics* 4: 1–27.
Goldin, C., and K. Sokoloff. 1982. Women, children, and industrialization in the early Republic: Evidence from the manufacturing censuses. *Journal of Economic History* 42: 741–774.
Graham, J., and C. Green. 1984. Estimating the parameters of a household production function with joint products. *Review of Economics and Statistics* 66: 277–282.
Gronau, R. 1973a. The effect of children on the housewife's value of time. *Journal of Political Economy* 81(Supplement): S168–S199.
Gronau, R. 1973b. The intrafamily allocation of time: The value of the housewives' time. *American Economic Review* 63: 634–651.
Gronau, R. 1973c. The measurement of output in the non-market sector – The evaluation of housewives' time. In *The measurement of economic and social performance*, ed. M. Moss, 163–199. New York: Columbia University Press.
Gronau, R. 1977. Leisure, production and work – The theory of the allocation of time revisited. *Journal of Political Economy* 85: 1099–1124.
Hartmann, H.I. 1981. The family as the locus of gender, class and political struggle: The example of housework. *Signs* 6: 366–394.
Hausman, J., and P. Ruud. 1984. Family labor supply with taxes. *American Economic Review* 74(2): 242–248.
Heckman, J. 1974. Effects of child care programs on women's work effort. *Journal of Political Economy* 82: 136–163.
Heckman, J. 1976. A life cycle model of earnings, learning and consumption. *Journal of Political Economy* 84(Supplement): S11–S44.
Heckman, J. 1978. A partial survey of recent research on the labor supply of women. *American Economic Review* 68(Supplement): 200–207.
Heckman, J. 1980. Sample selection bias as a specification error. In *Female labor supply*, ed. J. Smith, 206–248. Princeton: Princeton University Press.
Heckman, J., and T. MaCurdy. 1986. Labor econometrics. In *Handbook of econometrics*, vol. 3, ed. Z. Griliches and M. Intriligator. New York: North-Holland.
Heckman, J., M.R. Killingsworth, and T. MaCurdy. 1981. Empirical evidence on static labour supply models: A survey of recent developments. In *The economics of the labour market*, ed. Z. Hornstein, J. Grice, and A. Webb. London: HMSO.
Hicks, J.R. 1946. *Value and capital*, 2nd ed. Oxford: Oxford University Press.
Hicks, J.R. 1965. *The theory of wages*, 2nd ed. London: Macmillan.
Hill, M.A. 1983. Female labor force participation in developing and developed countries: Consideration of the informal sector. *Review of Economics and Statistics* 65: 459–468.
Hill, M.A. 1984. Female labor force participation in Japan: An aggregate model. *Journal of Human Resources* 19: 280–287.
Himmelweit, S., and S. Mohun. 1977. Domestic labour and capital. *Cambridge Journal of Economics* 1: 15–31.
Humphries, J. 1977. Class struggle and the persistence of the working-class family. *Cambridge Journal of Economics* 1: 241–258.
Joshi, H. 1985. Participation in paid work: Multiple regression analysis of the women and employment survey. In *Unemployment, job search and labour supply*, ed. R. Blundell and I. Walker, 217–246. Cambridge: Cambridge University Press.
Joshi, H., and S. Owen 1985. Does elastic retract? The effect of recession on women's labour force participation. Discussion Paper No. 64, Centre for Economic Policy Research, London.
Joshi, H., and S. Owen. 1984. How long is a piece of elastic? The measurement of female activity rates in British Censuses 1951–1981. Discussion Paper No. 31, Centre for Economic Policy Research, London.
Killingsworth, M.R. 1983. *Labor supply*. New York: Cambridge University Press.
Killingsworth, M.R. 1985. A simple structural model of heterogeneous preferences and compensating wage differentials. In *Unemployment, job search and labour supply*, ed. R. Blundell and I. Walker. Cambridge: Cambridge University Press.
Killingsworth, M.R., and J.J. Heckman. 1986. Female labor supply: A survey. In *Handbook of labor economics*, ed. O. Ashenfelter and R. Layard, 103–204. New York: North-Holland.
Kniesner, T. 1976. An indirect test of complementarity in a family labor supply model. *Econometrica* 44: 651–659.
Kooreman, P., and A. Kapteyn. 1984. *Estimation of rationed and unrationed household labor supply functions using flexible functional forms*, Research Memorandum, vol. 157. Tilburg: Department of Econometrics, Tilburg University.
Kooreman, P., and A. Kapteyn. 1985. *Estimation of a game theoretic model of household labor supply*, Research Memorandum, vol. 180. Tilburg: Department of Econometrics, Tilburg University.
Layard, R., and J. Mincer (eds.). 1985. *Trends in women's work, education, and family building*. Special edition of *Journal of Labor Economics* 3(1): S1–S396.
Leibowitz, A. 1974. Production within the household. *American Economic Review: Papers and Proceedings* 62(2): 243–250.
Leuthold, J. 1968. An empirical study of formula income transfers and the work decision of the poor. *Journal of Human Resources* 3: 312–323.
Long, C.D. 1958. *The labor force under changing income and employment*. Princeton: Princeton University Press.
Lundberg, S. 1985. The added worker effect. *Journal of Labor Economics* 3: 11–37.
Manser, M., and M. Brown. 1979. Bargaining analyses of household decisions. In *Women in the labor*

market, ed. C.B. Lloyd, E. Andrews, and C. Gilroy, 3–26. New York: Columbia University Press.
Manser, M., and M. Brown. 1980. Marriage and household decision-making: A bargaining analysis. *International Economic Review* 21: 31–44.
Marshall, A. 1920. *Principles of economics*, 8th ed. New York: Macmillan.
Martin, J., and C. Roberts. 1984. *Women and employment: A lifetime perspective*. London: HMSO.
McElroy, M., and M. Horney. 1981. Nash-bargained household decisions: Toward a generalization of the theory of demand. *International Economic Review* 22: 333–349.
Mincer, J. 1962. Labor force participation of married women: A study of labor supply. In *Aspects of labor economics*, ed. National Bureau of Economic Research, 63–97. Princeton: Princeton University Press.
Mincer, J. 1963. Market prices, opportunity costs and income effects. In *Measurement in economics*, ed. C.F. Christ, 67–82. Stanford: Stanford University Press.
Mincer, J. 1966. Labor force participation and unemployment: A review of recent evidence. In *Prosperity and unemployment*, ed. R.A. Gordon and M.S. Gordon, 73–112. New York: Wiley.
Mincer, J. 1985. Intercountry comparisons of labor force trends and of related developments: An overview. *Journal of Labor Economics* 3(Supplement): S1–S32.
Mincer, J., and H. Ofek. 1982. Interrupted work careers: Depreciation and restoration of human capital. *Journal of Human Resources* 17: 358–370.
Mincer, J., and S. Polachek. 1974. Family investments in human capital: Earnings of women. In *Economics of the family: Marriage, children and human capital*, ed. T.W. Schultz, 397–429. New York: Columbia University Press.
Mincer, J., and S. Polachek. 1978. Women's earnings reexamined. *Journal of Human Resources* 13: 118–134.
Moffitt, R., and K.C. Kehrer. 1981. The effect of tax and transfer programs on labor supply: The evidence from the income maintenance programs. In *Research in Labor Economics*, vol. 4, ed. R.G. Ehrenberg, 103–150. Greenwich: JAI Press.
Moffitt, R., and K.C. Kehrer. 1983. Correction. In *Research in labor economics*, vol. 6, ed. R.G. Ehrenberg, 452. Greenwich: JAI Press.
Mroz, T.A. (forthcoming). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Unpublished manuscript, Department of Economics, University of Chicago, Chicago. Forthcoming in *Econometrica*.
Nakamura, A., and M. Nakamura. 1981. A comparison of the labor force behavior of married women in the United States and Canada, with special attention to the impact of income taxes. *Econometrica* 49: 451–490.
Nakamura, A., M. Nakamura, and D. Cullen. 1979. Job opportunities, the offered wage, and the labor supply of married women. *American Economic Review* 69: 787–805.
Owen, J. 1986. *Working lives: The American work force since 1920*. Lexington: D.C. Heath.
Pencavel, J. 1986. The labor supply of men. In *Handbook of labor economics*, ed. O. Ashenfelter and R. Layard. New York: North-Holland.
Pigou, A.C. 1946. *The economics of welfare*, 4th ed. London: Macmillan.
Polachek, S. 1979. Occupational segregation among women: Theory, evidence and a prognosis. In *Women in the labor market*, ed. C.B. Lloyd, E.S. Andrews, and C.L. Gilroy, 137–157. New York: Columbia University Press.
Polachek, S. 1981. Occupational self-selection: A human capital approach to differences in occupational structure. *Review of Economics and Statistics* 63: 60–69.
Renaud, P.S.A., and J.J. Siegers. 1984. Income and substitution effects in family labour supply. *De Economist* 132: 350–366.
Robbins, L. 1930. Note on the elasticity of demand for income in terms of effort. *Economica* 10: 123–129.
Robinson, J.P. 1977. *Changes in Americans' use of time: 1965–1975, a progress report*. Cleveland: Communication Research Center, Cleveland State University.
Robinson, J.P., and P.E. Converse. 1967. *66 basic tables of time budget research data for the United States*. Ann Arbor: Survey Research Center, University of Michigan.
Robinson, C., and N. Tomes. 1985. More on the labour supply of Canadian women. *Canadian Journal of Economics* 18: 156–163.
Smith, R.E. (ed.). 1979. *The subtle revolution*. Washington, DC: Urban Institute.
Smith, S. 1983. Estimating annual hours of labor force activity. *Monthly Labor Review* 106(2): 13–22.
Smith, J.B., and M. Stelcner 1985. Labor supply of married women in Canada, 1980. Working Paper No. 1985–7, Department of Economics, Concordia University, Montreal.
Smith, J.P., and M. Ward 1984. *Women's wages and work in the twentieth century*. Report R–3119–HICHD. Santa Monica: Rand Corporation.
Smith, J.P., and M. Ward. 1985. Time-series growth in the female labor force. *Journal of Labor Economics* 3(Supplement): S59–S90.
Sorrentino, C. 1983. International comparisons of labor force participation, 1960–81. *Monthly Labor Review* 106(2): 23–36.
Stelcner, M., and J. Breslaw. 1985. Income taxes and the labor supply of married women in Quebec. *Southern Economic Journal* 51: 1053–1072.
Stelcner, M., and J.B. Smith 1985. Labour supply of married women in Canada: Non-convex budget constraints and the CES utility function. Working Paper No. 1985–9, Department of Economics, Concordia University, Montreal.
van der Veen, A., and G.H.M. Evers. 1984. A labour-supply function for females in the Netherlands. *De Economist* 132: 367–376.

Vanek, J. 1974. Time spent in housework. *Scientific American* 231(November): 116–120.
Wales, T., and A.D. Woodland. 1980. Sample selectivity and the estimation of labor supply functions. *International Economic Review* 21: 437–468.
Wales, T., and A.D. Woodland. 1983. Estimation of consumer demand systems with binding nonnegativity constraints. *Journal of Econometrics* 21: 263–285.
Yamada, T., and T. Yamada 1984. Part-time employment of married women and fertility in urban Japan. Working Paper No. 1474, National Bureau of Economic Research, Cambridge, MA.
Yamada, T., and T. Yamada 1985. Part-time work vs. full-time work of married women in Japan. Working Paper No. 1608, National Bureau of Economic Research, Cambridge, MA.
Yamada, T., T. Yamada, and F. Chaloupka 1985. A multinomial logistic approach to the labor force behavior of Japanese married women. Working Paper No. 1783, National Bureau of Economic Research, Cambridge, MA.

Labour Surplus Economies

Gustav Ranis

Abstract

In some sectors with a large endowment of unskilled labour and without sufficient cooperating land or capital, given technology and a wage level bounded from below, labour markets cannot clear. A full employment solution would drive remuneration below socially acceptable, possibly subsistence, levels of consumption. Consequently, a labour surplus exists in that much of the labour force contributes less to output than it requires: its marginal product falls below its remuneration, set by bargaining. A reallocation of such workers to other, competitive, sectors would eliminate the inefficiency and enhance total output. Open economy dimensions, extensions and critiques are dealt with.

Keywords

Agriculture and economic development; Arrow, K.; Balanced growth; Becker, G.; Behavioural economics; Dual economies; Engel's Law; Family networks; Informal sector; Kuznets, S.; Labour surplus; Marginal productivity; Minimum subsistence level of existence (MSL); Neoclassical economics; Population growth; Agricultural vs non-agricultural productivity; Sen, A.; Solidarity networks

JEL Classifications

O1

Labour surplus economies are closely associated with the concept of economic dualism, that is, the existence of organizational heterogeneity as between major sectors of an economy. The basic premise is that there exist some sectors or subsectors in which, in the presence of a large endowment of unskilled labour and the absence of sufficient cooperating land or capital, and with a given technology and a wage level bounded from below, labour markets cannot clear. A full employment, neoclassical 'wage equals marginal product' solution would drive remuneration below socially acceptable, possibly subsistence, levels of consumption. Consequently, a labour surplus exists in the sense that a substantial portion of the labour force contributes less to output than it requires, that is, its marginal product falls below its remuneration, set by bargaining. The 'labour surplus' designation then arises from the fact that a reallocation of such workers to other, competitive, or neoclassically functioning sectors would eliminate the aforementioned inefficiency and thus materially enhance the total output of the system.

The prime location for such surplus labour has traditionally been developing countries' agricultural sectors, concentrated especially in subsistence agriculture, characterized by family farms, that is, excluding commercialized plantation agriculture which consists of profit maximizing entities able to hire and fire workers following well-known neoclassical principles. Surplus labour makes its appearance in the context of owner-operated extended family networks, communes, villages or similar tenurial arrangements, all configurations in which income or output shares are determined via bargaining in relation to (though

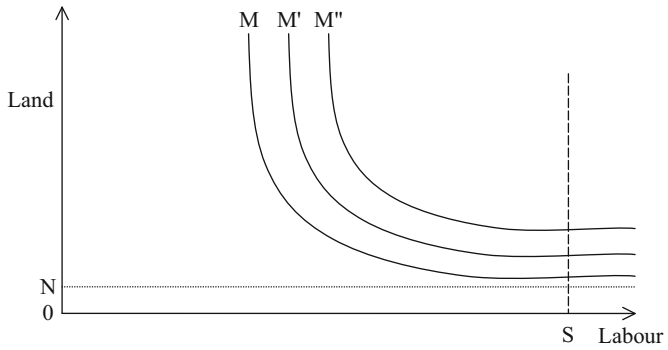
not necessarily equal to) the average rather than the marginal product of labour. Wage determination is thus based on a sharing principle, a function of the fact that, when high man-land ratios are among the initial conditions, low marginal-productivity workers cannot be dismissed or otherwise eliminated.

Here, we first present the static version of the labour surplus economy. Next we describe the conditions for balanced growth. Then open economy dimensions are introduced. Finally, some extensions are cited and rejoinders offered to some critiques.

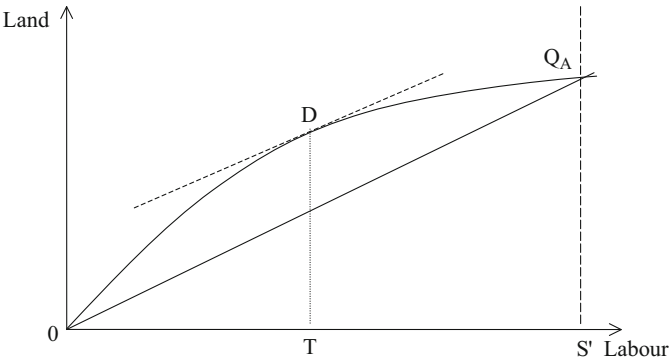
The Static Labour Surplus Economy

Figure 1 illustrates the situation of relatively scarce land, intensively cultivated, yielding extremely low increments of output at the margin. Labour is measured on the horizontal and land on the vertical axis, with production contour lines indexed as M, M', and M'' in Fig. 1. Given technology, fixed land at ON, and labour endowment at $OS = OS' = OS''$ in Figs. 1, 2 and 3, the total product curve is ODQ_A in Fig. 2 and the marginal product of labour, depicted by curve ABC in Fig. 3, approaches very low levels, substantially below the bargaining or institutional wage or income share OW_a which is related to (again, not necessarily equal to) the average product (slope of OQ_A in Fig. 2). Under these conditions, we can locate the proportion of the total agricultural labour force which is 'in surplus' in the sense that it is 'disguisedly unemployed' or 'underemployed' as $S''T$ in Fig. 3. This includes all those whose marginal product lies below their consumption or income share. They represent the 'labour surplus' phenomenon or what Rosenstein-Rodan (1943) and Nurkse (1953) long ago designated as 'hidden rural savings' which could be mobilized via reallocation to higher-productivity activities elsewhere in the economy.

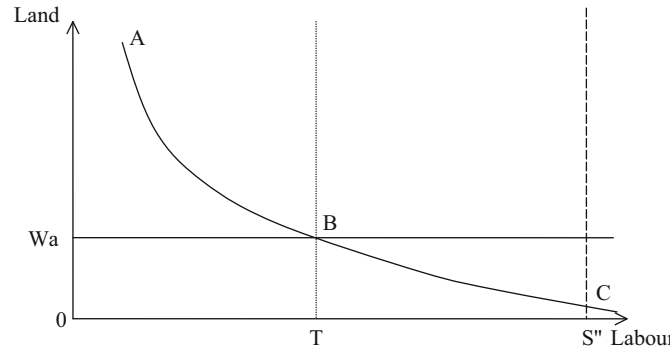
Labour Surplus Economies, Fig. 1



Labour Surplus Economies, Fig. 2



Labour Surplus Economies, Fig. 3



It should be emphasized that 'labour surplus' therefore does not mean, as has often been asserted, that a substantial portion of the agricultural labour force can be withdrawn without loss of output. Such a zero marginal product condition constitutes a statistically highly unlikely razor's edge event but, partly because it has been assumed for purely diagrammatic and/or mathematical convenience by Lewis (1972), by Fei and Ranis (1964), and by others, it has drawn extensive and often intemperate critical comment in the literature. Schultz (1964, p. 70), for example, cited the fact that output in India fell with a decline in the agricultural working population due to an influenza epidemic as proof that surplus labour was a 'false doctrine'. As Sen (1967) pointed out in rebutting Schultz on this point, when some workers with low (or even zero) marginal productivity are withdrawn, some of those left behind are likely to adjust by working harder. Or, put more broadly, any withdrawal of labour from agriculture is very likely to be accompanied by a reorganization of production arrangements on the part of those left behind, that is, by technology change. This would be equivalent to an upward shift of the ODQ_A curve in Fig. 2 and of the *ABC* curve in Fig. 3.

Balanced Growth in the Labour Surplus Economy

Dynamically, the labour surplus condition can thus be seen as permitting an increasing number of agricultural workers and an increasing volume of agricultural surplus, defined as the difference between total agricultural output and what is needed to satisfy the remaining agricultural population's consumption requirements, to move out and support the expansion of commercialized activities, industry and services, rural and urban. This labour surplus condition of the economy then ultimately comes to an end when increases in agricultural productivity, which free up workers and generate agricultural surpluses and, accompanied by increases in productivity in the expanding commercialized sector, enhance the demand for workers, have proceeded in a more or less 'balanced' fashion long enough, and at a rate exceeding population growth, to mop up the disguisedly unemployed, that is, all those whose marginal product lies below their wage or consumption standard.

This critical concept of the need for 'balance' between the non-commercialized and commercialized components of the labour surplus economy has really three ingredients. One, the most obvious, is that the release of labour from non-commercialized agriculture is roughly in balance with its absorption by commercialized non-agriculture. Another, focused on the product rather than the organizational dimension of dualism, suggests that relative advances in productivity in the two sectors proceed in such a fashion that the inter-sectoral terms of trade are not substantially affected, that is, that the system does not encounter food shortages or, less likely, food surpluses in the course of the development process.

Third, the financial intermediation network, primitive at first, more sophisticated later, represents a crucial link as it must be capable of transforming non-commercialized sector surpluses, joined by commercialized sector profits, into efficient investment, mainly in the commercialized sector.

To turn first to more specifics on the inter-sectoral labour market, it should be noted that the unskilled real wage in the commercialized sector will tend to be tied to, though certainly not equal to, the non-commercialized agricultural real wage. A substantial unskilled labour wage gap is indeed likely to be required, partly to induce the typical agricultural worker to overcome her attachment to soil and family, partly to meet transport costs, and partly as a consequence of such institutional factors as commercialized sector minimum wage legislation, unionization, the public sector wage setting, and so forth, all of which usually do not extend into non-commercialized activities. Once these two wage levels are given within a general equilibrium context, the release of labour by the non-commercialized sector and its absorption by the commercialized sector represents an essential ingredient of balanced growth in the labour surplus economy.

It should also be noted that both wages may be expected to rise over time, in part because, as agricultural sector labour productivity increases, there is also likely to be some upward adjustment of the bargaining wage which is tied to the rising average product. Moreover, the inter-sectoral wage gap may rise as a consequence of a change in the extent of commercialized sector interventions via minimum wage increases, enhanced union bargaining power, and so on. The two unskilled real wage patterns over time may thus be conceived of as a step function, horizontal at any point in time, reflecting the labour surplus condition, but at a slightly higher level, again horizontal, in the next period. All this will, of course, yield a gently rising labour supply curve over time, giving way to a sharply rising pattern once the labour surplus has been exhausted and remuneration is determined neoclassically, that is to say, by the marginal product. Meanwhile, the existence of a relatively constant or gently upward-sloping real wage over time in both sectors, with a possibly growing gap between them, can be expected to induce labour-intensive technology choices and, more importantly, labour-using technological change in both the non- commercialized and commercialized sectors of the labour surplus economy.

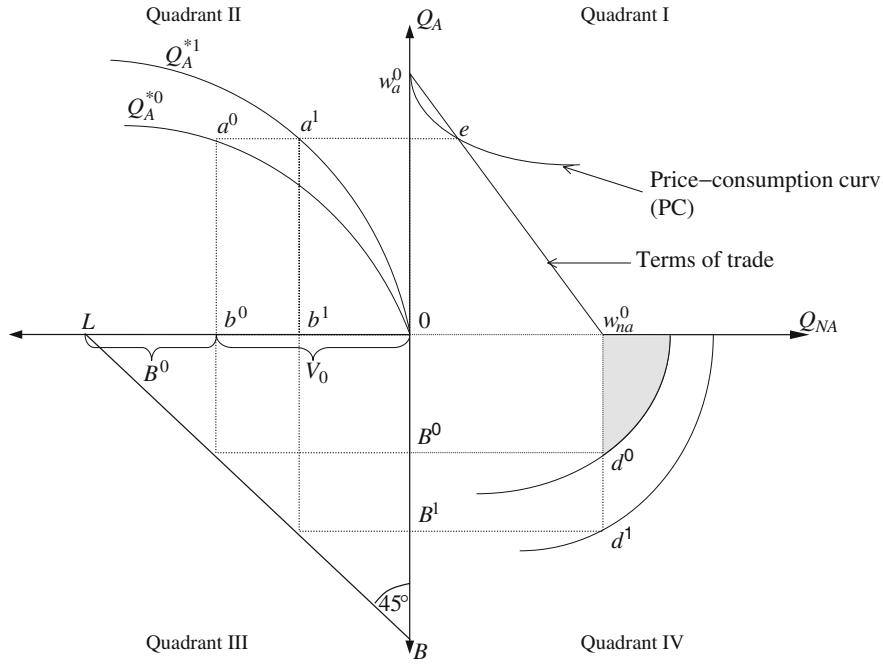
Second, an understanding of the workings of the inter-sectoral commodity market is required for an assessment of the contribution of the non-commercialized sector to the rest of the economy. This can be seen in terms of the net real resources transferred, that is, the difference between the shipments of food and raw materials delivered to the commercialized sector and the shipments of goods and services sent in the opposite direction. The agricultural sector's export surplus may thus be viewed as the contribution of that sector to both the labour reallocation and overall growth process over time.

The main participants in the dualistic commodity market are thus, on the one hand, the owners of the agricultural surplus and, on the other, the newly allocated workers who may be thought of as receiving wage income in the form of non-agricultural goods and anxious to trade some of these for the food 'left behind'. Once this transaction is completed, the reallocated worker finds herself in possession of the agricultural goods needed to at least maintain her consumption standard – most likely to increase it because of the aforementioned inter-sectoral wage gap. In this fashion the dualistic commodity market is indispensable for transforming the consumption bundle of the agricultural labour force into a wages fund for the newly allocated non-agricultural workers. At the same time the owners of the agricultural surplus, such as the landlords and/or the government via land taxes, obtain a claim against a portion of the newly formed non-agricultural capital stock; the other portion results from the reinvestment of profits by commercialized sector entrepreneurs. The above underlines the importance of the product, along with the organizational dimension of balanced growth in the closed labour surplus economy, rooted in the fact that food and non-agricultural products cannot readily be substituted for each other. Agriculture is thus a necessary condition for non-

agriculture, while the converse does not strictly hold. In the open economy, food imports, of course, become possible, thus helping the system avoid premature food shortages, as illustrated by Japan's historical experience in the early decades of the 20th century (see Hayami and Ruttan 1970).

Third, the financial counterpart of the real resources contribution of the non- commercialized to the commercialized sector over time is effected through the workings of the intersectoral financial market. As we have seen, the savings of the agricultural sector become a claim against non-agriculture, the magnitude of which is determined by the size of its export surplus. These savings must somehow be channeled into non-agricultural investment; that is, what is left of the agricultural surplus that is not siphoned off by consumption or intermediate input requirements must find its way into capital formation in the rest of the economy.

The dynamics reflecting all the main facets of such a balanced growth path can be illustrated by reference to Fig. 4 within a simplified setting, that is, without intermediate input flows between the two sectors. Total population L is shown on the horizontal axis in quadrant II, moving from right to left, with agricultural output and the institutional consumption standard $c = w_a^0$ measured in terms of agricultural goods, on the vertical axis. The curve OQ^{*0}_A describes per capita food availability for the total population, or Q/L, at a given level of technology, for various possible proportions, θ, of the total population already allocated to other activities, B, that is, ($\theta = B/L \geq 0$). One equilibrium point along a balanced growth path may then be defined as follows: let initial consumption $c = w_a^0$, and the terms of trade between w_a^0, the 'wage in terms of agricultural goods' (Q_A), and w_{na}^0 the 'wage in terms of non-agricultural goods' (Q_{NA}) be given. For simplification only, we assume that there is no wage gap between unskilled agricultural and non-agricultural workers. The price–consumption curve (PC) in quadrant I of Fig. 4 then indicates all possible points of tangency between changing terms of trade and a given typical worker's consumer preference between agricultural and non-agricultural goods. Point e is the consumption equilibrium



Labour Surplus Economies, Fig. 4

point for the typical worker, given the terms of trade shown, regardless of whether she is engaged in agricultural or non-agricultural activities. B is the population outside of agriculture and the remaining agricultural population V ($L = B + V$) produces enough food to meet everyone's consumption requirements at the institutional wage.

The auxiliary 45° line in quadrant III transposes workers B^0, already allocated to non-agricultural work, onto the vertical axis, that is, OB^0. The consistent equilibrium point for employment in the non-agricultural sector is then point d^0, located at the intersection between the 'horizontal' supply curve of non-agricultural labour, at wage level w^0_{na}, and the demand curve for non-agricultural labour, or the marginal productivity curve corresponding to a particular level of the capital stock and technology in that sector. This describes an equilibrium position $a^0b^0d^0$ in both the intersectoral labour and commodity markets.

To turn to the definition of balanced growth over time, and on the assumption of no upward adjustment of the agricultural real wage and, thus, of the non- agricultural real wage which is 'tied' to it, balanced increases in agricultural and non-agricultural productivity resulting from capital accumulation and technology change can be shown by a shift of the per capita food availability curve to OQ^*_A1 in quadrant II, with Lb^1 or OB^1 workers now allocated, as well as of the marginal productivity of non-agricultural labour curve to d^1 in quadrant IV. This would result in a new equilibrium position $a^1b^1d^1$ where, once again, the two intersectoral markets clear. Such a growth path would clearly meet the labour market equilibrium condition, and a little more work would permit us to demonstrate that equilibrium in the commodity market sense, as previously defined, also continues to be achieved, permitting agricultural and non-agricultural workers to exchange some of the goods they produce for the goods they need, at the given terms of trade, enabling everyone to remain at the same equilibrium point e.

To turn to the inter-sectoral financial market, the landlords and/or the government, whoever owns the agricultural surplus, would end up with a claim against some part of the non-agricultural capital stock. This, plus the reinvested industrial profits represented by the shaded area in quadrant IV of Fig. 4 would be invested in the non-agricultural sector, causing, along with technology change, the indicated shift of the marginal productivity curve. The investment fund for the next period is thus composed of this period's savings out of the agricultural surplus plus the savings out of non-agricultural profits. For the sake of convenience, we have made the assumption of no leakage into consumption by either landlords or capitalists. The allocation of the society's investment fund plus its innovative energies, as between the sectors, would then be guided by the relative shortages of agricultural and non-agricultural goods, as reflected, in the case of a market economy, by changes in the inter-sectoral terms of trade. In a non-market economy the role of changes in the terms of trade as a signalling device would be taken over by evidence of unplanned shortages or surpluses in the material balances sense. We have here again made a simplifying, but not critical, assumption that technology change is responsible for agricultural productivity change, while all the investment funds are allocated to non-agriculture.

As we have already noted, the entire transition process must not only be balanced but also proceed at a pace in excess of population growth if the initial reservoir of surplus labour is to ultimately be exhausted and neoclassical wage determination is to take over. Moreover, if balanced growth, as indexed by the rate of labour reallocation, only marginally exceeds the rate of population growth on average, the length of time it takes to arrive at the commercialization point, marking the end of labour surplus, must also be politically acceptable.

The real world, of course, does not quite operate in such a smooth fashion. There are times when, under the impetus of an 'industry first' strategy, non- agricultural productivity increases for some time at a rate in excess of agricultural productivity growth, leading to food shortages, the shifting of the terms of trade in favour of agriculture, and an increase in the non-agricultural real wage. The reverse can also occur, although

empirically there seems to be less danger of that. Most successful labour surplus societies (such as historical Japan and post-war South Korea, Taiwan and Thailand) have, in fact, experienced something approaching constancy in the terms of trade.

In any case, progress along a balanced growth path at a rate in excess of population growth – and sufficiently in excess to guarantee a politically acceptable time perspective – is essential to a society's successful transition into a modern growth regime. Success is defined as the end of labour surplus, that is, the end of organizational dualism in the labour market. Once balanced growth has proceeded long enough and fast enough labour surplus gives way to labour shortage in both sectors, which means that the marginal productivity calculus of wage determination takes over. At this point organizational dualism disappears; and, given considerable increases in per capita incomes and the workings of Engel's Law, product dualism also atrophies over time as agriculture gradually becomes an appendage to the economy, or just another symmetrical sector within the system's input–output matrix. Increasingly the economy is then ready to perform according to the rules of modern economic growth as described by Simon Kuznets (1966).

Open Economy Dimensions

Thus far we have discussed the development of the labour surplus economy mainly in a closed economy context. The open economy or trade-related dimensions of development in the labour surplus economy are, of course, important enough to warrant substantial amendment of the analysis presented here. During the early colonial, or open agrarian, phase of development, the economy may well be tied to foreign markets by virtue of some of the labour force being weaned away from food production and into land-based export-oriented activity: for example, minerals and other primary products of interest to foreign investors. This typically leads to a triangular relationship among the cash-crop export sector, the foreign sector, and the food producing domestic agricultural sector. But once the economy moves out of its colonial or 'overseas territory' phase and into a national development-oriented effort, our analysis must be amended to take 'openness' into account.

To do so, we must, first, recognize that the export-oriented cash crop agricultural sub-sector continues to generate foreign exchange earnings but that these are now used, in addition to possible food imports, to assist in the construction of a new, domestically oriented, non-agricultural sector producing previously imported non-durable consumer goods, that is, to fuel so-called primary or 'easy' industrial import substitution. These raw material-intensive exports thus provide a second source of agricultural surplus which, converted into industrial capital goods imports, and possibly supplemented by the inflow of foreign savings, helps finance non- agricultural growth in the same balanced growth context. In this way a new triangular relationship between two kinds of commercialized activities, one agricultural and the other non-agricultural, plus the food producing non-commercialized agricultural hinterland, replaces the colonial triangle.

What happens at the end of this primary import substitution phase is critical; that is, once domestic markets for the non-durable consumer goods are exhausted, it is apparent that relatively natural resources rich labour surplus countries have a tendency to continue with import substitution, now shifting from labour-intensive light industries to the more capital-intensive durable consumer goods, the processing of raw materials, and the production of capital goods. At the same time, in the minority of countries which have a relatively poor natural resources base we observe a shift from a domestic to an export-market orientation for the same labour-intensive non-durable consumer goods. In that case the export sector now constitutes a powerful new production function available to the economy through which traditional and, later, non-traditional exports can be converted into imported capital goods and raw materials. Moreover, the openness of the economy permits foreign capital to provide additional finance in support of the balanced growth process. Finally, an important potential advantage of the economy's openness is, of course, the whole

range of additional technological alternatives now made available, which, hopefully with modifications and adaptations, can help increase the efficiency and speed of the balanced growth process.

The open economy, in other words, not only permits the labour surplus economy to harvest the normal gains from trade, to benefit from the vent for surplus of previously underutilized resources – in this case not only raw materials but also unskilled labour – but also, dynamically, to affect the direction of technology change and thus introduce competitive forces and ideas from abroad which are able to diffuse throughout the economy and are undoubtedly of considerable importance in determining the success of the labour surplus economy's transition efforts.

Extensions and Critiques

Up to now we have focused exclusively on owner-operated agriculture as the typical representative of the non-commercialized sector of the labour surplus economy. It should, however, be recognized that there are very likely to exist substantial portions of non-agricultural activities, both rural and urban, and both industry and services- oriented, which are labour surplus in the way we have defined the condition. This time, the cooperating factor in short supply is capital. Most relevant is the so-called informal sector – both rural but most heavily urban – which occupies a large, often dominant, position in many developing countries. Family and cooperative ventures in this setting are characterized by the same sharing of total income, that is, a bargaining wage, coupled with low marginal productivity, that we encountered in subsistence agriculture. We are here including not only the substantial portions of both the rural and urban populations engaged in distributive trades and services – ranging from the vendors of tea, flowers and cigarettes to barbers, bootblacks and car watchers – but also to blacksmiths, metal workers, and repair shops that dominate the landscape in most labour surplus developing countries. Some portions of this informal sector, especially its urban branch, are likely to be static and of the labour- absorptive 'sponge' variety; others may be capable of technology change, of subcontracting arrangements with the urban formal or commercialized sector as well as of generating surpluses for investment in that sector. Thus, organizational dualism is quite pervasive in both rural and urban non-agriculture, even as product dualism now loses its distinctive characteristic.

As development since the 1950s has proceeded apace, some initially labour surplus countries, including Taiwan, South Korea and Thailand, have graduated from their initial labour surplus condition, evidenced by gently rising unskilled wages in both sectors, finally giving way to rapid and sustained increases as secular labour shortages make their appearance. Such a turning point was reached around 1968 in the case of Taiwan, around 1973 in the case of South Korea and around 1993 in the case of Thailand. It is also true that many developing countries, starting with up to 80 per cent of their population and 50 per cent of their output in food producing agriculture, have gradually shifted substantially into non-agricultural pursuits, with services retaining their dominant position, even as their composition has changed radically, in the commercialized direction. As a consequence, the number of contemporary developing countries with typical initial labour surplus characteristics has been declining. Nevertheless, a large preponderance of the developing world, certainly by weight of population, continues to find itself in a labour surplus condition. This holds, for example, for China and India, huge countries both currently engaged in a vigorous balanced growth effort, as well as for other parts of South Asia, much of Central America, the Caribbean and parts of South America. Even some countries of sub-Saharan Africa, once considered land surplus by some observers, may, as a consequence of population growth and the loss of land to the Sahara, be approaching labour surplus status – though, given the AIDS epidemic, this remains a more controversial issue.

It should, finally, be noted that the fundamental concept of the labour surplus economy has come under increasing attack by the dominant neoclassical school of economics. While still viewed as relevant in the South and wherever

heavy population pressure on scarce cultivable land remains a feature of the landscape, most Northern economists in the Becker micro-econometric tradition find it difficult to accept the notion of an exogenous or bargaining wage in the non-commercialized sectors instead of one determined endogenously by the customary interaction between demand and supply. The crux of the critique is based on the rejection of the notion that initial conditions, that is, a highly unfavourable ratio of people to cooperating land or capital, can lead to the subsidization of some members of the society by others, in lieu of ejecting them.

The work of Rosenzweig and associates (for example, Rosenzweig 1988), presenting evidence of rising labour supply curves in a cross-section of such heavily populated agricultural sectors as India's, typifies current mainstream rejection of the 'unlimited supply of labour' condition underlying the labour surplus economy construct. Yet we would contend that such efforts capture an expressly static snapshot picture, addressing cross-sectional labour–leisure decisions across households already working at full capacity (that is, with little leisure to spare), while labour surplus models are concerned with the conditions governing inter-sectoral labour reallocation over time.

The exogenous agricultural wage assumption underpinning labour surplus economies, so troubling to neoclassical economists, gets support from anthropologists like Geertz (1963) and Scott (1976), as well as from economists like Lewis (1972), Ishikawa (1975), Fei and Ranis (1964), Osmani (1991), Ohkawa (1972) and others. Fafchamps (1992) provides an overview of the principles underlying the 'solidarity network' among peasants as depicted in anthropological evidence. Ishikawa (1975), long an astute observer of Asian economic development, endorses the concept of a 'minimum subsistence level of existence' (MSL), one version of the institutional real wage. His work indicates the prevalence of a 'community principle of employment and income distribution'. This principle promises all member MSL families. . . an income not less than MSL' (Ishikawa 1975, p. 474). Hayami and Kikuchi (1982, p. 217), basically neoclassical in outlook, find that in Indonesia

> . . . wage rates cannot adjust directly to changes in labor's marginal productivity. Adjustments in wage rates are allowed only through modification of institutional arrangements themselves . . . In other words, 'institutional wages' based on a system of community-wide work and income-sharing similar to the classical concept can adjust to the neoclassical equilibrium through institutional innovations.

Only over time is there a tendency to adjust, but even then it does not necessarily occur by altering wages to equal the marginal product, which could reduce the wage below subsistence. Instead, in Java harvest contracts are adjusted to include weeding duties without a complementary rise in the wage rate, thereby not threatening the MSL but moving institutionally towards equilibrium. Even Kenneth Arrow (1988), one of the high priests of neoclassical economics, states that it may take a considerable period of time before equilibrium is reached. Osmani (1991) presents a model of downward rigidity of the sharing rule insisted on by the workers themselves. Current work in what is called behavioural economics may also prove to be of help in developing a theoretical structure to rationalize cross-worker subsidization in the absence of assured reciprocity – especially as some members of the group are likely to be leaving agriculture over time.

Perhaps even more relevant, there is evidence, not only for Taiwan, Korea, and Thailand but also for post-enclosure England between 1780 and 1840 and for post- Restoration Japan between 1870 and 1920, indicating substantial increases in agricultural labour productivity while both agricultural and non-agricultural unskilled real wages were rising only gently, until commercialization was reached and wages began to rise steeply in line with rising marginal productivity. Thus, both historical and 20th-century development patterns are inconsistent with the neoclassical school's one-sector full-employment equilibrium assumptions.

In the final analysis, what is relevant is whether the labour surplus model provides a better fit for the observed empirical pattern of successful labour-abundant developing countries; whether

the model is better suited to analysing relative agricultural neglect in failure cases; whether it is better able to explain changing patterns of technology choice and the direction of technology change; whether, in sum, it makes better sense than to assume away the initial existence of underemployment and disequilibrium before the one-sector, fully commercialized modern growth epoch can be reached.

See Also

- ▶ Agriculture and Economic Development
- ▶ Classical Growth Models
- ▶ Dual Economies

Bibliography

Arrow, K. 1988. General economic theory and the emergence of theories of economic development. Presidential address, 8th World Economic Congress of the International Economic Association. New Delhi, 1986.

Fafchamps, M. 1992. Solidarity networks in preindustrial societies: Rational peasants with a moral economy. *Economic Development and Cultural Change* 41: 147–174.

Fei, J., and G. Ranis. 1964. *Development of the labor surplus economy: Theory and policy*. Homewood: Richard A. Irwin, Inc.

Geertz, C. 1963. *Agricultural involution: The process of ecological change in Indonesia*. Berkeley: University of California Press.

Hayami, Y., and M. Kikuchi. 1982. *Asian village economy at the crossroads*. Baltimore: Johns Hopkins University Press.

Hayami, Y., and V. Ruttan. 1970. Korean rice, Taiwan rice, and Japanese agricultural stagnation: An economic consequence of colonialism. *Quarterly Journal of Economics* 84: 562–589.

Ishikawa, S. 1975. Peasant families and the agrarian community in the process of economic development. In *Agriculture in development theory*, ed. L. Reynolds. New Haven: Yale University Press.

Kuznets, S. 1966. *Modern economic growth: Rate, structure and spread*. New Haven: Yale University Press.

Lewis, W. 1972. Reflections on unlimited labor. In *International economics and development: Essays in honor of Raul Prebisch*, ed. L. DiMarco. New York: Academic Press.

Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. New York: Oxford University Press.

Ohkawa, K. 1972. *Differential structure and agriculture: Essays on dualistic growth*. Tokyo: Kinokuniya.

Osmani, S.R. 1991. Wage determination in rural labor markets: The theory of implicit cooperation. *Journal of Development* 34: 3–23.

Rosenstein-Rodan, P. 1943. The problem of industrialization of eastern and south-eastern Europe. *Economic Journal* 53: 202–211.

Rosenzweig, M. 1988. Labor markets in low income countries. In *Handbook of development economics*, ed. H. Chenery and T.N. Srinivasan, vol. 1. Amsterdam: North Holland.

Schultz, T. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.

Scott, J.C. 1976. *The moral economy of the peasant*. New Haven: Yale University Press.

Sen, A. 1967. Surplus labor in India: A critique of Schultz' statistical test. *Economic Journal* 77: 154–161.

Labour Theory of Value

Fernando Vianello

JEL Classifications
B1

The only instance in which Adam Smith makes the value of commodities depend on the quantity of labour required to produce them is where 'the whole produce of labour belongs to the labourer' (Smith 1776, vol. 1, p. 54; see ibid., p. 72). 'In that early and rude state of society which precedes both the accumulation of stock and the appropriation of land', he asserts 'the proportion between the quantities of labour necessary for acquiring different objects seems to be the only circumstance which can afford any rule for exchanging them for one another' (ibid., p. 53).

This contention is illustrated by the famous example of the beaver and the deer:

> If among a nation of hunters, for example, it usually costs twice the labour to kill a beaver which it does to kill a deer, one beaver should naturally exchange for or be worth two deer. It is natural that what is usually the produce of two days or two hours labour, should be worth double of what is usually the produce of one day's or one hour's labour. (ibid., p. 53)

According to Smith, when profit and rent make their appearance alongside the labourer's income, the above rule is no longer applicable. The price of

a commodity is then obtained by adding up its 'component parts': wage, profit and rent. These revenues, which Smith calls 'the three original sources ... of all exchangeable value' (ibid., p. 59), enter into the 'natural price' of each commodity at their respective 'natural rates', such that 'the natural price itself varies with the natural rate of each of its component parts, of wages, profit and rent' (vol. I, p. 71).

The 'adding-up' theory of prices must be distinguished from Smith's claim that the price of every commodity 'resolves itself' entirely into wage, profit and rent (see vol. I, p. 57). The latter was accepted by Ricardo and rejected by Marx. The former was rejected by both.

1. Against the 'adding-up' theory Ricardo sets the labour theory of value extended to the capitalist mode of production:

> All the implements necessary to kill the beaver and deer might belong to one class of men, and the labour employed in their destruction might be furnished by another class; still their comparative prices would be in proportion to the actual labour bestowed, both on the formation of the capital, and on the destruction of the animals. (Ricardo 1821, p. 24)

The value of the product would go partly to the labourers and partly to the capitalists; yet this division could not affect the relative value of these commodities, since whether the profits of capital were greater or less, whether they were 50, 20 or 10 per cent or whether the wages of labour were high or low, they would operate equally on both employments. (ibid.)

As gold, the standard of value, is a commodity like any other, the above argument makes the price of commodities – the exchange-ratio between each of them and gold – independent of the level of the wage, a change in which is exactly offset by a change in the opposite direction of the rate of profits: the relative weight of the two 'component parts', wages and profits, varies, but their sum remains the same.

According to Ricardo the value of a commodity produced from natural resources in short supply is regulated by the quantity of labour expended to produce it 'under the most unfavourable circumstances ... under which the quantity of produce required, renders it necessary to carry on the production' (ibid., p. 73). Thus the quantity of labour governing the value of the entire quantity produced of a commodity is not that actually expended on its production, but that which would need to be expended if the entire production took place under the most unfavourable circumstances. That portion of the value which is absorbed by rent corresponds to the difference between this fictitious quantity of labour and the one actually expended on the production of the commodity. The portion of value corresponding to the quantity of labour actually expended is split up into wages and profits.

Thus the labour theory of value enables Ricardo to conceive the different revenues as resulting from the breakdown of a known magnitude, rather than that magnitude (value) as resulting from the adding up of 'component parts' (the different revenues) determined independently of each other. The contrast between these two conceptions is fixed by Marx in a highly effective image:

> If I determine the lengths of three different straight lines independently, and then form out of these three lines as 'component parts' a fourth straight line equal to their sum, it is by no means the same procedure as when I have some given straight line before me and for some purpose divide it, 'resolve' it, so to say, into three different parts. In the first case, the length of the line changes throughout with the lengths of the three lines whose sum it is; in the second case, the lengths of the three parts of the line are from the outset limited by the fact that they are parts of a line of given length. (Marx 1885, p. 387)

2. If gold is produced by an unchanging quantity of labour, a rise in the price of a commodity can only stem from a process of 'extensive' or 'intensive' diminishing returns (only the former, however, will be considered in what follows). In discussing the consequences of an increasing 'difficulty of procuring the necessaries on which wages are expended', Ricardo takes the quantities consumed by each labourer as given. It follows that, as the price of corn (a typical necessary) rises, the wage in terms of gold also rises, and the profits of the manufacturers fall:

> suppose corn to rise in price because more labour is necessary to produce it; that cause will not raise the

> prices of manufactured goods in the production of which no additional quantity of labour is required. If, then, wages continued the same, the profits of manufacturers would remain the same; but if, as is absolutely certain, wages should rise with the rise of corn, then their profits would necessarily fall. (Ricardo 1821, pp. 48, 110–11)

Let us assume that the entire production of corn is initially obtained from land of uniform quality, and that thereafter, in order to increase the quantity produced, land of an inferior quality be brought into cultivation. The value of the quantity of corn produced on the second quality of land is governed by the quantity of labour actually expended on its production and 'is divided into two portions only: one constitutes the profits of stock, the other the wages of labour' (ibid., p. 110). The increase in the value of the quantity of corn obtained from the first quality of land is wholly swallowed up by the rent, which now begins to be paid for the use of this quality of land.

In the production of corn both expenses and proceeds per unit of produce increase. But the result is the same as in manufacturing (where only expenses increase) since the farmer 'will not only have to pay, in common with the manufacturer, an increase of wages to each labourer he employs, but he will be obliged either to pay rent, or to employ an additional number of labourers to obtain the same produce; and the rise in the price of raw produce will be proportioned only to that rent, or that additional number, and will not compensate him for the rise of wages' (ibid., p. 111).

What causes the ratio of profits to wages to fall is not the rise of rent, but – in agriculture as well as in manufacturing – the increase in wages consequent upon the increased expenditure of labour required to produce necessaries in the most unfavourable circumstances. If the commodities which increase in value are not among those purchased by labourers, the ratio of profits to wages remains unchanged (even though a part of the capitalist's purchasing power is transferred to the landowners).

3. What is true of the ratio of profits to wages is also true, in Ricardo's opinion, of the rate of profits, which forms his main concern. Indeed, what he does is simply to refer to the latter his conclusions regarding the former, so that the two concepts appear to shade into one another. 'In his observations on profit and wages', says Marx, taking up a remark of G. Ramsay's (1836, p. 174n.), 'Ricardo . . . treats the matter as though the entire capital were laid out directly in wages' (Marx 1905–10, vol. II, p. 373). Marx traces this confusion back to 'the absurd dogma pervading political economy since Adam Smith, that in the final analysis the value of commodities resolves itself completely into . . . wages, profit and rent' (Marx 1894, p. 841).

Smith's teaching is that, while the price of a commodity includes – along with the revenues derived from its direct production – the value of its means of production, the latter value can be broken down in the same way, and so on, going backwards, until an *initial stage of production* is reached, in which the means of production of the stage following are produced without the aid of any other means of production. Only the value of the output in the initial stage of production resolves itself immediately into wage, profit and rent. But the output in each stage, whose value equals the sum of the revenues obtained in that stage as well as in all the preceding ones, supplies the means of production for the next stage, so that 'the whole price still resolves itself either immediately or ultimately into the same three parts of rent, labour, and profit' (Smith 1776, vol. I, p. 57; here 'labour' obviously stands for 'wages').

Marx's criticism of Smith's thesis of complete 'resolution' of prices into revenues is made up of two parts, which should be kept strictly distinct. The first is of a factual nature. In moving back from a commodity to its means of production, and from these to their own means of production, and so on, one will never – in Marx's view – reach an initial stage of production, since sooner or later one is bound to encounter commodities that, either directly or indirectly, participate in the production of themselves. Since one can never get rid of these commodities, however far back one goes, 'it is [of] no avail for Adam Smith to send us from pillar to post' (Marx 1905–10, vol. I, p. 99).

The conception according to which commodities are produced in a finite number of stages does not, of itself, lead to a confusion between the rate

of profits and the ratio of profits to wages. Since, however, in this conception the value of the means of production employed in each stage resolves itself into the revenues obtained in all the previous stages, 'one may ... imagine along with Adam Smith' – this being the second part of Marx's criticism – 'that constant capital is but an apparent element of commodity-value, which disappears in the total pattern' (Marx 1894, p. 845; by 'constant capital' Marx means the value of the means of production).

That in dealing with the economy as a whole Smith and Ricardo fall into this error emerges clearly, for example, from Smith's statement, repeated almost *verbatim* by Ricardo, according to which 'what is annually saved is as regularly consumed as what is annually spent, and nearly in the same time too; but it is consumed by a different set of people' (Smith 1776, vol. I, p. 359; see Ricardo 1821, p. 151n.). The funds devoted to accumulation are here treated as wholly employed in producing the necessaries for the labourers. This may help explaining how, when Ricardo approaches the problem from the point of view of the economy as a whole, he does not seem to make any distinction between the rate of profits and the ratio of profits to wages, referring to the former as depending only on the 'proportion of the annual labour of the country [which] is devoted to the support of the labourers' (Ricardo 1821, p. 49; see Sraffa 1951, p. xxxiii).

4. Although it is the labour theory of value that makes it possible for Ricardo to determine the rate of profits, his adherence to this theory appears anything but firm. Indeed, 'the principle that the quantity of labour bestowed on the production of commodities regulates their relative value' turns out to be, as Ricardo puts it, 'considerably modified' (ibid., p. 30) by the influence of other factors.

To show this Ricardo makes use of a numerical example which deserves to be quoted in full:

> Suppose I employ twenty men at an expense of £ 1,000 for a year in the production of a commodity, and at the end of the year I employ twenty men again for another year, at a further expense of £1,000 in finishing or perfecting the same commodity, and that I bring it to market at the end of two years, if profits be 10 per cent., my commodity must sell for £2,310; for I have employed £1,000 capital for one year, and £2,100 capital for one year more. Another man employs precisely the same quantity of labour, but he employs it all in the first year; he employs forty men at an expense of £2,000, and at the end of the first year he sells it with 10 per cent. profit, or for £2,200. Here then are two commodities having precisely the same quantity of labour bestowed on them, one of which sells for £2,310 – the other for £2,200. (ibid., p. 37)

Let w be the wage (equal in the example to £50 per labourer) and r be the rate of profits (equal to 10 per cent). For the sake of simplicity, we shall further suppose that the quantity produced of each of the two commodities be one unit. The price of commodity a, the first commodity in the example, is then

$$20w(1+r)^2 + 20w(1+r) = P_a$$

The price of the second commodity, b, is instead

$$40w(1+r) = P_b$$

Although Ricardo does not deal systematically with the subject, here, as well as in other numerical examples, he does offer a theory in embryo, which – for any given rate of profits – makes natural prices depend not only on the quantity of labour directly or indirectly expended on each commodity, but also on what we may call the *distribution over time* of that quantity of labour.

5. Since in the foregoing example the prices of the two commodities are determined on the basis of prior knowledge of the wage and the rate of profits, one may be inclined to think, with Marshall, that according to Ricardo value is regulated by the cost of production, which includes 'Time or Waiting as well as Labour'; and that Marx wrongly interpreted his doctrine 'to mean that interest does not enter into that cost of production which governs ... value' (Marshall 1920, p. 672 and pp. 672–3, n. 1). That this is not the case will emerge clearly if we look at Ricardo's approach to the problem of relative price variation as set forth in a numerical example contained in his 1823 paper on *Absolute Value and Exchangeable Value* (Ricardo 1823, pp. 383–4); an example which closely follows the one we have just examined (the only differences, which we shall ignore,

being that the prices of commodities *a* and *b* corresponding to $r = 10$ per cent are said to be £231 and £220 respectively, rather than £2,310 and £2,200, and that a third commodity is also considered).

Ricardo supposes 'labour to rise in value and profits to fall – that from 10 pc^t they fall to 5 pc^t. He further supposes that commodity *b* be the standard of value. Making the two examples into a single one, we shall suppose that gold is produced in a single stage. If, then, the price of commodity *b* is £2,200, that is not because the wage is £50 and the rate of profits 10 per cent, but rather because it has been produced, like gold, in a single stage, employing a quantity of labour equal to 2,200 times that required to produce the quantity of gold corresponding to £1. The fall in the rate of profits from 10 per cent to 5 per cent will thus leave the price of commodity *b* unchanged; which amounts to saying that in its production (as in that of gold) the increase in wages and the fall in profits offset each other.

However, the same increase in *w* and fall in *r* cannot bring about a similar offsetting in the case of commodity *a*, whose price must fall from £2,310 to £2,255 (from £231 to £225.5 in Ricardo's 1823 example). This result is obtained by applying the rate of profits of 5 per cent (instead of 10 per cent) to the value of the means of production employed in the second stage of production of commodity *a*. The latter value, £1,100, does not vary, since the means of production are produced, like gold (and commodity *b*), in a single stage. The value of the term $20w(1 + r)^2$ in the equation of commodity *a* falls, therefore, from £1,210 to £1,155. The value of the second term in the sum, $20w(1 + r) =$ £1,100, can be assimilated to the unchanging value of a commodity produced in a single stage.

It is evident that, if gold were produced in two years, with the same proportional distribution of labour between the two corresponding stages of production as commodity *a,* the new ratio P_a/P_b would emerge from a rise in P_b with P_a constant. It is also evident that, if all commodities were produced with the same proportional distribution of labour over time, they would all be in the same situation as gold, in whose production an increase (fall) in wages is exactly offset by the corresponding fall (increase) in profits, and the labour theory of value would stand in no need of 'modification'.

The 'modifications' have, therefore, nothing to do with the alleged necessity of adding to the labour what is depicted as a second element of the cost of production. The misunderstanding may be traced back to Malthus, who ascribes to Ricardo the very fault that Marshall seeks to acquit him of, shifting the blame onto Marx. 'We have the power indeed', Malthus remarks:

> arbitrarily, to call the labour which has been employed upon a commodity its real value, but in so doing, we use words in a different sense from that in which they are customarily used; we confound at once the very important distinction between *cost* and *value*; and render it almost impossible to explain with clearness, the main stimulus to the production of wealth, which in fact depends upon that distinction.

To which Ricardo counters:

> Mr Malthus appears to think that it is part of my doctrine, that the cost and value of a thing should be the same; – it is, if he means by cost, 'cost of production' including profits. In the above passage, this is what he does not mean, and therefore he has not clearly understood me. (Ricardo 1821, p. 47n.)

What Ricardo makes clear in this passage (which, surprisingly enough, Marshall quotes as evidence in support of his reading of the matter: see Marshall 1920, p. 672) is that the labour theory of value, in its 'unmodified' as well as its 'modified' form, takes full account of 'the very important distinction between cost and value'; that is, of the existence of profits ('the main stimulus to the production of wealth'). What equals value according to this theory is not, Ricardo argues, 'cost' as commonly understood, but 'cost of production including profits', profits being what is left of the value of a commodity once wages have been deducted. (Reference to the most unfavourable circumstances under which production is carried on has been dropped since the preceding section, land being now supposed to be abundant and all of the same quality.)

6. The reader will perhaps have noted how Ricardo omits to specify by how much the wage must increase in order to cause a fall from ten to

five per cent in the rate of profits (elsewhere, again when dealing with the problem of relative price variation, he postulates 'such a rise of wages as should occasion a fall of one per cent. in profits': Ricardo 1821, p. 36). Even though Ricardo continues to express himself as if, in the relation between w and r, the independent variable were represented by the wage, in actual fact he reverses the roles, and makes w depend on r. The value of w when $r = 10$ per cent is, as we know, $w =$ £50. Its value when $r = 5$ per cent can be calculated from the equation of commodity b (whose price remains £2,200). This value is slightly less than $w =$ £52.8 s. 0d.

As a matter of fact, Ricardo's argument is made up of two distinct stages. In the first of these the rate of profits is determined on the basis of the 'unmodified' labour theory of value; in this stage the necessaries consumed by each labourer are taken as given (see Sect. 2 above). The second stage takes the rate of profits as given, the problem being now to determine the prices which make the rate of profits uniform throughout the economy. These prices, as Ricardo realizes, are not regulated by the quantities of labour expended on the production of the commodities, as they were assumed to be for the purpose of determining the rate of profits. And the wage (the £52.8 s. 0d or so of the example) will in general turn out to be different from the value of the necessaries it was assumed to purchase in the first stage of the argument.

It does not escape Ricardo that the rate of profits should be determined on the basis of the 'modified' theory, and therefore of prices which, in turn, cannot be determined before the rate of profits is known. But he is unable to provide a theoretical construction capable of coping with this interdependence. Thus he does not see any other solution but that of continuing to base his analysis of income distribution on the 'unmodified' labour theory of value, which he defends as 'the nearest approximation to truth as a rule for measuring relative value, as any I have ever heard' (Letter to Malthus of 9 October 1820, in Ricardo 1951–73, vol. VIII, p. 279).

7. A major difference between the Ricardian version of the labour theory of value and its Marxian version, to which we must now turn, lies precisely here: that the former can be described as an approximation, whereas the latter cannot. According to Marx the values of commodities exactly (not approximate) reflect the quantities of labour expended on their production, although this is not true, in general, of the 'prices of production' (Marx's name for 'natural prices'), which coexist with values.

In discussing Marx's position we shall reckon the value of commodities directly in units of labour (say, man-years). The value of the means of production which assist one labourer in the annual cycle of production of any particular commodity, or *constant capital* per unit of labour (c), and the value of one labourer's necessaries, or *variable capital* per unit of labour (v), are thus made equal to the quantities of labour expended on the production of those means of production and of those necessaries respectively.

If only circulating capital is used, the value of the output per unit of labour of any commodity is $(c + 1)$, or c plus the value added per unit of labour. Since v is uniform throughout the economy (each labourer being assumed to consume the same bundle of commodities), the *surplus-value* per unit of labour $(1 - v)$ will also be uniform. The same is obviously true of the ratio of surplus-value to variable capital (the *rate of surplus–value),* but not, in general, of the ratio of surplus–value to total (i.e. constant plus variable) capital. The latter ratio will be the higher, in any particular branch of production, the lower is the ratio c/v (the *organic composition of capital*).

Competition, however, redistributes the overall surplus-value of the economy among the various branches of production in such a way as to render it proportional not to the variable, but to the total capital. Thus a general rate of profits comes to be established, equal to the weighted average of the $(1 - v)$ to $(c + v)$ ratios in the different branches of production – or, which amounts to the same thing, to the ratio of the overall surplus-value of the economy to the overall capital employed. The same mechanism establishes the prices of production, which make that rate of profits uniform throughout the economy.

Unlike Ricardo's Marx's argument is *explicitly* framed in two stages. Since the prices of production

differ from the values only on account of the different distribution of the overall surplus-value of 'the economy, according to Marx the rate of profits is accurately determined, for the economy as a whole, on the basis of the labour theory of value. The prices of production are then obtained from the values by replacing the surplus-value produced in each branch of production with the part of the overall surplus-value of the economy belonging to that branch according to the general rate of profits.

8. 'Surplus-value and the rate of surplus-value', says Marx, 'are, relatively, the invisible and unknown essence that wants investigating, while rate of profit and therefore the appearance of surplus-value in the form of profit are revealed on the surface of the phenomenon' (Marx 1894, p. 43). To reveal the invisible: herein lies the task of science. But Marx's theoretical programme also involves explaining just *why* the intimate essence of things in invisible, why it does not reveal itself 'on the surface of the phenomenon'. Marx's explanation is that those 'who are entrapped in bourgeois production relations' (ibid., p. 817) witness the *result* of the redistribution of surplus-value–the profit proportional to capital – but not the *process* leading up to this result:

> The actual difference of magnitude between profit and surplus-value . . . in the various spheres of production now completely conceals the true nature and origin of profit not only from the capitalist, who has a special interest in deceiving himself on this score, but also from the labourer. (ibid., p. 168)

Thus it comes about that

> the splitting of the value of commodities after subtracting the value of the means of production consumed in their creation; the splitting of this given quantity of value, determined by the quantity of labour incorporated in the produced commodities into three component parts . . . appears in a perverted form on the surface of capitalist production,

wage, profit and rent taking on the aspect of 'independent revenues in relation to one another, and as such related to three very dissimilar production factors, namely labour, capital and land', from which 'they seem to arise' (ibid., pp. 867–8; we shall, however, continue to assume the absence of rent). 'To have destroyed this false appearance and illusion' represents 'the great merit of [classical] political economy' (ibid., p. 830). Against classical political economy – of which Ricardo is the 'last great representative' (Marx 1873, p. 24) – Marx sets 'vulgar' economy: the first of these studied 'the real relation of production in bourgeois society', whereas the second 'deals with appearances only' (Marx 1867, p. 85, n. 1).

But even Ricardo cannot be completely acquitted, in Marx's opinion, of having taken as the *starting-point* of the argument the *result* of the redistribution of surplus-value. Indeed, it is the natural prices themselves that Ricardo claims are regulated (even if only approximately; but, as will be remembered, it is the nearest approximation to truth' among those available; see Sect. 6 above) by the quantities of labour expended on the production of commodities. Hence Marx's allegation that Ricardo confuses values and prices of production.

If Ricardo is compelled to presuppose what he should explain (the profit proportional to capital, as it emerges from the redistribution of surplus-value), this is – according to Marx – because his unsatisfactory treatment of non-wage capital (see Sect. 3 above) blinds him to the distinction between surplus-value and profit:

> Ricardo wrongly identifies surplus-value with profit . . . these are only identical in so far as the total capital consists of variable capital or is laid out directly in wages . . . Ricardo evidently shares Smith's view that the *total value* of the annual product resolves itself into revenues. Hence also his confusion of value with cost-price. (Marx 1905–10, vol. II, p. 426; as so often in *Theories of Surplus-Value,* 'cost-price' here stands for 'price of production')

Here, in Marx's opinion, lies the origin of the analytical difficulties with which Ricardo had to wrestle and which Marx himself claims to have overcome, thanks to his discovery of the redistribution mechanism.

9. On 24 August 1867, a few days after correcting the proofs of the first volume of *Capital,* Marx wrote to Engels:

> The best points in my book are: (1) the *double character of labour,* according to whether it is expressed in use value or exchange value (*all* understanding of the facts depends upon this . . .) (2) the

> treatment of *surplus-value independently of its particular forms* as profit, interest, ground rent, etc. (Marx and Engels 1942, pp. 226–7)

The second of these two contributions has been dealt with in Sects. 7 and 8 above (and something more on the subject will be said in Sect. 11 below), within the limits of the hypothesis that all surplus-value is received in the form of profit. We must now turn to the first contribution – the one on which *'all* understanding of the facts' is based: the 'double character of labour'.

In the production of commodities the distribution of labour in a society among its various productive activities is not regulated *a priori*, through some form of agreement or coercion, but only *a posteriori*, through the exchange of products (Marx 1867, p. 336). The labour of individuals is therefore not, immediately, the labour *of society* – as is the case in, say, a peasant family, within which 'the labour-power of each individual, by its very nature, operates . . . merely as a definite portion of the whole labour-power of the family' (Marx 1867, p. 82; see Marx 1859, p. 33). On the contrary, we are dealing here with 'the labour of private individuals or groups of individuals who carry on their work independently of each other'; this labour 'asserts itself as a part of the labour of society, only by means of the relation which the act of exchange establishes directly between the products, and indirectly, through them, between the producers' (Marx 1867, pp. 77–8). It is only when the social division of labour takes this particular form that the products of labour become commodities, or acquire the quality of possessing value.

In the first chapter of *Capital* (as well as in the first chapter of *A Contribution to the Critique of Political Economy*) Marx emphasizes how in the eyes of producers commodities count not for their ability to satisfy this or that human want, but rather for their ability to find a purchaser: not for their use-value but for their (exchange-) value. Of these two qualities of commodities, use-value is the one abstracted from in the exchange, which cancels the difference between the products, in the sense that in the exchange different products are equated, or treated as equal, and reduced to their quality of possessing value.

Labour participates in the two-fold character of commodities, as useful things and things possessing value. On the one hand, 'it must, as a definite useful kind of labour, satisfy a definite social want, and thus hold its place as a part and parcel of the collective labour of all, as a branch of a social division of labour' (Marx 1867, p. 78). On the other hand, just as 'in viewing the coat and linen as values, we abstract from their different use-values, so it is with the labour represented by those values: we disregard the difference between its useful forms, weaving and tailoring' (ibid., p. 52); which is what producers themselves actually do, production of commodities being *production for value* – production, therefore, of abstract wealth, indifferent to its material content. What remains is a uniform, undifferentiated labour, which 'counts only quantitatively', having been 'reduced to human labour, pure and simple' (p. 52), to 'abstract human labour' (p. 81). Such is the labour which, embodied in commodities, figures as their value.

'Whenever, by an exchange', Marx writes, 'we equate as values our different products, by that very act, we also equate, as human labour, the different kinds of labour expended upon them. We are not aware of this, nevertheless we do it' (ibid., pp. 78–9). The reduction of a commodity to its mere quality of possessing value and the reduction of labour to abstract labour are thus in Marx's conception the outcome of one and the same real process (see Colletti 1968, Sect. 8). And it is only by being reduced to abstract labour and assuming the form of a quality of commodities, their value, that the *private* labour of the weaver and the *private* labour of the tailor enter into relation with each other, becoming part of a *social* division of labour. This is, in Marx's words, 'the specific manner in which the social character of labour is established' (Marx 1859, p. 32) in the production of commodities. 'But what is the value of a commodity?', Marx enquires. 'The objective form of the social labour expended on its production' (Marx 1867, p. 501). Or, to put it another way, abstract labour (social only in so far as abstract) represents 'the substance of value' (ibid., p. 46).

10. The picture is now complete, and we can attempt to gather together the threads of Marx's

position. As we have just seen, the thesis of the reduction of labour to abstract labour is put forward by Marx in close connection with his theory of value. Indeed, the two merge into one, abstract labour being indicated as the substance of value and value as the form that labour must assume in order to acquire a social character. It remains to be added that the conception of abstract labour as the substance of value presupposes the sort of redistribution mechanism described in Sect. 7 above. What constitutes the substance of value cannot, in fact, but constitute the substance of revenues, as the latter stem from the breakdown of the value of a given set of commodities. It follows that the conception of abstract labour as the substance of value necessitates that the whole of this substance be found in the prices of production, having merely been partly diverted away from some commodities and channelled into others (see the enlightening comparison with the 'conservation of energy' in Lippi 1976, pp. 50–52). If this is not the case, then the aforesaid substance is not the 'substance' of anything real, and 'value' is merely a name for the quantity of labour directly and indirectly expended on the production of a commodity.

11. In the Afterword to the second (German) edition of *Capital* we read that 'the method of presentation must differ in form from that of inquiry' (Marx 1873, p. 28). We are, now in a position to understand this celebrated (as much as hermetic) warning. If we attend to the 'method of inquiry', the theory of the rate of profits and of the prices of production (contained in the manuscripts published posthumously as the third volume of *Capital*) represents – as stated in the preceding section – a premise for the conception of abstract labour as the substance of value, and the cornerstone of the whole theoretical structure of *Capital*. (From a chronological point of view, it has been remarked that 'once Marx had attained – at the beginning of 1858 – what he regarded as the correct solution of the problem of how to determine the rate of profit, various elements in his thinking seem to have found an organic unity in the concept of value – the concept of a "substance" to be redistributed' (Ginzburg 1985, pp. 105–6); the 'various elements' being basically Marx's analysis of the social division of labour and his theory of income distribution and prices.)

But if, instead, we attend to the 'method of presentation', things take on a rather different aspect. Marx calls his own presentation of the argument 'genetical', meaning by this that it consists in 'elaborating how the various forms come into being' (Marx 1905–10, vol. III, p. 500), proceeding from the form of value that labour assumes in the act of acquiring a social character, to arrive at surplus-value, the redistribution mechanism and the establishment of a general rate of profits.

The two 'methods', or procedures, reflect the two different aims mentioned in Sect. 8 above: the aim (proper to scientific analysis) of tearing away the veil of appearances, and the aim (proper to genetical presentation) of showing how that veil is woven together. The latter aim is not regarded by Marx as less important than the former, to explain how appearances are produced being in his opinion the only sure way of evading their deceptions.

As we have already seen, Ricardo himself is believed by Marx to be partly the victim of such deceptions, even while he contributed so greatly towards dispelling them. In conceiving the labour theory of value as a theory of natural prices, Ricardo 'omits some essential links and *directly* seeks to prove the congruity of economic categories with one another' (Marx 1905–10, vol. II, p. 165). He does so by taking 'the rate of profits as something pre-existent which, therefore, even plays a part in the determination of *value'* (ibid., p. 434), thus missing the inner connection of forms which is reflected in Marx's genetical presentation, and according to which 'the determination of value is the primary factor, antecedent to the rate of profits and to the establishment of production prices' (ibid., vol. III, p. 377; see Gajano 1979, ch. 3).

12. If, however, the presentation must proceed from value to the rate of profits and the prices of production, it must assume (at least provisionally) that the foundation of value be independent of what comes after, as a result of the redistribution of surplus-value. Marx thus finds himself in an *impasse,* no such independent foundation being provided by his analysis.

So it comes as no surprise that value is introduced in *Capital* in a rather sketchy way. Marx starts by declaring, as something self-evident, that in two commodities equated in exchange 'there exists in equal quantities something common to both' (Marx 1867, p. 45). He then goes on to enquire wherein this common element consists. It is at this point that we meet the argument according to which exchange involves an abstraction from the use-value of the commodities exchanged ('the exchange of commodities is evidently an act characterised by a total abstraction from use-value': (ibid., p. 45; see Sect. 9 above). But, Marx pursues, 'if then we leave out of consideration the use-value of commodities, they have only one common property left, that of being products of labour' (ibid., p. 45). Thus he does his best to lead the reader into thinking that the prices of commodities are regulated by the quantity of labour expended on their production (otherwise the common element would not be 'in equal quantities'). Only later on does Marx put the reader on his guard with sporadic and obscure hints. ('Average prices do not directly coincide with the values of commodities, as Adam Smith, Ricardo, and others believe': ibid., p. 163n.; see ibid., p. 212n., where the reader is referred to vol. III – unpublished – and ibid., p. 290, where Marx mentions the 'many intermediate terms' wanted to resolve the 'apparent contradiction' between the labour theory of value and the existence of a uniform rate of profits.)

'Analysis', writes Marx, 'is the necessary prerequisite of genetical presentation' (Marx 1905–10, vol. III, p. 500). But it is a prerequisite which cannot be openly declared if presentation is to remain genetical.

This limitation has given birth to two opposite and equally wrong interpretations. The one holds that Marx's theory of value has no foundation whatsoever, and treats that theory and the theory of prices of production as two mutually incompatible theories of prices (this is the thesis of the 'contradiction' between the first and the third volumes of *Capital,* put forward in Böhm-Bawerk 1896). The other interpretation tries to defend the labour theory of value on the basis of Marx's analysis of the social division of labour, making no appeal to the redistribution mechanism and maintaining, in the last analysis, that labour forms the substance of value because it is through the exchange of commodities that the various labours, performed outside any conscious coordination, enter into relation with one another (this traditional Marxist reply to Böhm-Bawerk's criticism first appears in Hilferding 1904, and finds its best expression in Colletti 1968).

Obviously the labour theory of value cannot be defended on the grounds indicated by Hilferding and Colletti (as the latter has acknowledged: see Colletti 1979). But, just as obviously, Böhm-Bawerk's grounds for dismissing it are not good ones. Actually, the reason why the labour theory of value must be rejected is not that it is devoid of foundation, but rather that what in Marx's view represents its foundation – his theory of the rate of profits and of prices of production – proves untenable in the light of the subsequent work of Tugan-Baranovsky (1905), Bortkiewicz (1907) and others, up to Sraffa (1960).

See Also

- British Classical Economics
- Marxian Value Analysis
- Natural Price
- Ricardo, David (1772–1823)

Bibliography

Colletti, L. 1968. Bernstein and the Marxism of the second international. In *From Rousseau to Lenin: Studies in ideology and society*, ed. L. Colletti. London: New Left Books. 1972.

Colletti, L. 1979. *Tra marxismo e no*. Bari: Laterza.

Gajano, A. 1979. *La dialettica della merce. Introduzione allo studio di 'Per la critica dell'economia politica' di Marx*. Napoli: il Laboratorio.

Ginzburg, A. 1985. A journey to Manchester: A change in Marx's economic conceptions. *Political Economy* 1.

Hilferding, R. 1904. Böhm-Bawerk's criticism of Marx. In Sweezy (1949).

Lippi, M. 1976. *Value and naturalism in Marx*. London: New Left Books. 1979.

Marshall, A. 1920. *Principles of economics,* 8th ed. London: Macmillan. 1964.

Marx, K. 1859. *A contribution to the critique of political economy*. Moscow: Progress Publishers. 1978.

Marx, K. 1867. *Capital: A critique of political economy,* vol. I. London: Lawrence & Wishart. 1977.
Marx, K. 1873. Afterword to the 2nd German edition. In Marx (1867).
Marx, K. 1885. *Capital: A critique of political economy,* vol. II. London: Lawrence & Wishart. 1974.
Marx, K. 1894. *Capital: A critique of political economy,* vol. III. London: Lawrence & Wishart. 1974.
Marx, K. 1905–10. *Theories of surplus-value.* Vol. I, 1978; Vol. II, 1975; Vol. III. Moscow: Progress Publishers, 1975.
Marx, K., and F. Engels. 1942. *Selected correspondence 1846–1895*. New York: International Publishers.
Ramsay, G. 1836. *An essay on the distribution of wealth*. Edinburgh: Adam & Charles Black.
Ricardo, D. 1821. On the principles of political economy and taxation. In Ricardo (1951–73), Vol. I. 3rd ed.
Ricardo, D. 1823. Absolute value and exchangeable value. In Ricardo (1951–73), vol. IV.
Ricardo, D. 1951–73. The works and correspondence of David Ricardo. Ed. P. Sraffa with the collaboration of M.H. Dobb. Cambridge: Cambridge University Press.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen. 1961.
Sraffa, P. 1951. Introduction. In Ricardo (1951–73), Vol. I.
Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.
Sweezy, P.M., ed. 1949. *Karl Marx and the close of his system by Eugen von Böhm-Bawerk & 'Böhm-Bawerk's Criticism of Marx' by Rudolf Hilferding, together with an appendix consisting of an article by Ladislaus von Bortkiewicz on the transformation of values into prices of production in the Marxian system*. New York: Augustus M. Kelley, 1966.
Tugan-Baranovsky, M. 1905. *Theoretische Grundlagen des Marxismus*. Leipzig: Duncker & Humblot.
von Böhm-Bawerk, E. 1896. Karl Marx and the close of his system. In Sweezy (1949).
von Bortkiewicz, L. 1907. On the correction of Marx's fundamental theoretical construction. In Sweezy (1949).

Labour's Share of Income

Douglas Gollin

Abstract

Economists have long studied labour's share of national income as a crude indicator of income distribution. More recently, labour's share has also been seen as offering insights into the shape of the aggregate production function. This has made labour's share a parameter of interest for macroeconomics, growth economics, and international economics, among other fields. Recent studies support the longstanding observation that labour's share of national income is relatively constant over time and across countries. Measurement of labour income, however, can be difficult in economies where many people are self-employed or work in family enterprises.

Keywords

Aggregation; Balanced growth; Cobb–Douglas functions; Constant-returns production function; Entrepreneurial income; Factor shares; Labour's share of income; National income accounting

JEL Classifications

D4; J10

At least since the time of Adam Smith, economists have been interested in the shares of production accruing to the owners of different factors. In the era before formalized national income and product accounts, factor shares were observed primarily at the firm or industry level. But Smith himself recognized that national product could similarly be divided into the income received by owners of land, labour and capital (the last of which he termed 'stock'). Early in Book I of *The Wealth of Nations*, Smith (1776, p. 155) notes that

> the exchangeable value . . . of all the commodities which compose the whole annual produce of the labour of every country, taken complexly, must resolve itself into . . . three parts and be parcelled out among different inhabitants of the country, either as the wages of their labour, the profits of their stock, or the rent of their land . . . Wages, profit, and rent, are the three original sources of all revenue as well as of all exchangeable value.

Smith and other early economists viewed the distribution of income among factors of production as intimately related to the level of wages and the degree of income inequality within a country. This was probably a reasonable assumption, given that, outside of agriculture and certain types of

self-employment, most individuals probably subsisted entirely on wage income.

Factor shares were, in fact, one of the few available sources of data on the size distribution of income – a subject that was viewed as crucial for policymaking, but about which little was known. As late as 1912, a prominent US labour economist wrote (Streightoff 1912, p. 155), 'Knowledge of the distribution of incomes is vital to sane legislative direction of progress. In a form definite enough for practical use, this knowledge does not exist. No time should be wasted in obtaining this knowledge.'

Labour's share of national income was seen as a particularly sensitive issue – intimately related to the supposed struggle of labour against capital. Simon Kuznets (1933, p. 30) referred to '[t]he significant political and social conflicts that center about the relative share of these productive factors'. Because of the importance of the topic, and because factor shares could be estimated reasonably well from micro data, a considerable literature emerged to document cross-section and time series observations on factor shares. In fact, the literature on factor shares eventually served as one of the foundations for the emergence of national income and product accounts.

From the beginning, the measurement of factor shares has been complicated by the difficulty of disentangling individual incomes into their functional components. Certain categories of income are easily assigned to land, labour, or capital. For example, wages and salaries are generally classifiable as labour income – although for some high-skill workers (such as hedge fund managers, star athletes), they may also embody some rents. Dividends and interest must be forms of capital income. Land rents are easily classified. But Kuznets (1933) pointed out that entrepreneurial income – which was about one fourth of national income in the 1920s – represented a mix of wages, salaries, interest, rent, and profits.

As national income accounting evolved over the succeeding decades, there were few improvements to the categorization of income according to factors of production. Irving Kravis (1962, p. 122) noted that 'the theory of distribution remains in a parlous state', largely because 'the components of income for which we have data has not been determined by the requirements of the economists but by the legal and institutional arrangements of our society'.

Nevertheless, by the 1950s a striking empirical regularity had begun to emerge. Labour's share of national income in the United States appeared to have remained roughly constant over a long period of time. Modest increases in the share of wages and salaries in national income appeared to have come at the expense of declines in entrepreneurial income – consistent with a structural shift away from self-employment and towards wage work. The regularity was sufficiently pronounced that Charles Cobb and Paul Douglas, writing in 1928, suggested that a simple constant-returns production function in the now familiar form $Y = AK^{1/4}L^{3/4}$ would provide an accurate representation of the US time series for aggregate output as a function of aggregate capital stock and labour. They considered a value for labour's share as low as two-thirds to be plausible.

As national income accounting became more systematic, evidence on factor shares accumulated over succeeding decades. John Maynard Keynes, writing in 1939 (p. 48), referred to the 'stability of the proportion of the national dividend accruing to labour, irrespective apparently of the level of output as a whole and of the phase of the trade cycle'. He went on to refer to this (p. 48) as 'one of the most surprising, yet best-established facts in the whole range of economic statistics, both for Great Britain and for the United States'.

D. Gale Johnson (1954) constructed and analysed data for the US economy going back over a century, to 1850, and concluded (p. 175) that there had been no 'significant secular change' in labour's share of income over that period. Robert Solow's paper (1957) on the sources of growth in the US economy noted that the data for the US economy seemed consistent with a Cobb–Douglas representation for the aggregate production function, with a capital share of 0.35 (and thus, implicitly, a labour share of 0.65). (However, Solow 1958, professed scepticism over the proposition that factor shares were actually constant, suggesting instead that variation within sectors was balanced out at the aggregate

level.) Nicholas Kaldor (1961) characterized the phenomenon as one of the stylized facts of modern economic growth.

This apparent consensus soon began to unravel, however. A major challenge to the hypothesis of constant factor shares appeared in comparisons of factor shares across countries. Kuznets, in an influential 1959 paper, further argued that the cross-country evidence did not support the view that factor shares were constant across countries or over time. Kuznets argued that data for other countries – and in particular for poor countries – revealed very different levels for labour's share in other countries. In particular, Kuznets suggested that labour's share of income was systematically lower in poor countries than in rich countries, while the share of unincorporated enterprises in national income was higher in poor countries than in rich countries. Kuznets concluded that the concept of a labour share lacked useful meaning – particularly as a proxy for discussions of the size distribution of income. His scepticism over constant factor shares was echoed by Solow (1958) and by Kravis (1962), among others.

To a large degree, scholarly interest in the labour share waned in succeeding years, although quantitative studies in both international trade and growth continued to rely on Cobb–Douglas aggregate production functions. In the trade literature, it was commonplace to assume that rich countries had a relatively high labour share, while poor countries had lower shares. Macro and growth studies of advanced economies typically assumed a Cobb–Douglas production function with a labour share of about two-thirds, often based on the employee compensation share of GNP for the United States, but this parametrization was seen as problematic for models that were intended to characterize both poor countries and rich ones.

This apparent discrepancy between cross-country and time series observations on labour's share was largely unaddressed in the literature until Gollin (2002) revisited the question. Drawing on the earlier work of Kuznets and others, he noted the potential significance of self-employment in skewing 'naive' calculations of factor shares. Gollin argued that poor countries typically have far higher levels of self-employment than do rich countries; as a result, cross-country comparisons of the employee compensation share (or wage share) will tend to yield large differences between rich and poor countries. Gollin showed that, after adjusting labour's share to account for differences in self-employment rates, no systematic patterns remained in the cross-country data between a country's income and its imputed labour share. Gollin reported labour shares in most countries, adjusted for self-employment, between 0.6 and 0.8. Similar results were obtained by Ben Bernanke and Refet Gürkaynak (2002), who used a different approach to adjust for the fraction of output produced by unincorporated enterprises.

Recent and preliminary work by Rodrigo García-Verdú (2005) for Mexico found that labour's share falls into this range when estimated from household survey data, rather than from national income accounts might suggest. However, Daniel

Ortega and Francisco Rodríguez (2006) present evidence from industrial census data that labour shares are lower in poor countries than in rich countries. And Samuel Bentolila and Gilles Saint-Paul (2003) show that labour's share within OECD countries is not constant, but rather moves in parallel with changes in the capital–output ratio.

Econometric studies of aggregate production functions, such as those by John Duffy and Chris Papageorgiou (2000) and Pol Antràs (2004), often reject the Cobb–Douglas specification of the aggregate production function. This suggests that, if factor shares are indeed (approximately) constant, there must be a different underlying mechanism. At the simplest level, any constant returns production function with labour-augmenting technical progress can give rise to constant factor shares if the rate of return on capital is constant over time – as, for example, on a balanced growth path. To see this, consider a simple Solow model with the constant returns aggregate production function $Y = F(K, AL)$. The productivity parameter A grows at a constant rate g, and there is an exogenous savings rate, s.

This economy will converge to a balanced growth path; assuming no population growth, the condition for balanced growth is given by

$$k^* = \frac{sf(k^*)}{\delta + g},$$

where δ is the depreciation rate and

$$k \equiv \frac{K}{AL}.$$

But the balanced growth path implies that the capital share is

$$\frac{rk^*}{f(k^*)} = \frac{sr}{\delta + g},$$

which will necessarily be constant because the rate of return is constant along the balanced growth path.

An alternative way to generate constant factor shares is through aggregation. Charles I. Jones (2005) reproduces and generalizes a result of Houthakker (1955) in which an aggregate Cobb–Douglas technology can be derived from firm-level or industry-level Leontief techniques. Jones shows that the same intuition can be applied more generally to a world in which the underlying production technologies have almost any form, and the 'aggregation' can simply occur across ideas or techniques within a firm. Jones's result is consistent with factor shares that are constant, but it also allows for movement in the factor shares and for differences across countries. In general, it appears to offer a useful theoretical framework for reconciling the different features of the data.

See Also

- ▶ Cobb–Douglas Functions
- ▶ Economic Growth
- ▶ Economic Growth, Empirical Regularities in
- ▶ Factor Prices in General Equilibrium
- ▶ Growth Accounting
- ▶ Level Accounting
- ▶ Production Functions

Bibliography

Antràs, P. 2004. Is the US aggregate production function Cobb-Douglas? New estimates of the elasticity of substitution. *Contributions to Macroeconomics* 4(1), Article 4.

Bentolila, S. and G. Saint-Paul. 2003. Explaining movements in the labor share. *Contributions to Macroeconomics* 3(1), Article 9.

Bernanke, B.S., and R.S. Gürkaynak. 2002. Is growth exogenous? Taking Mankiw, Romer and Weil seriously. In *NBER macroeconomics annual 2001*, ed. B.S. Bernanke and K.S. Rogoff. Cambridge, MA: MIT Press.

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

Duffy, J., and C. Papageorgiou. 2000. A cross-country empirical investigation of the aggregate production function specification. *Journal of Economic Growth* 5: 87–120.

García-Verdú, R. 2005. Factor shares from household survey data. Working paper no. 2005–05, Dirección General de Investigación Económica, Banco de México.

Gollin, D. 2002. Getting income shares right. *Journal of Political Economy* 110: 458–474.

Houthakker, H.S. 1955. The Pareto distribution and the Cobb–Douglas production function in activity analysis. *Review of Economic Studies* 23: 27–31.

Johnson, D.G. 1954. The functional distribution of income in the United States, 1850–1952. *Review of Economics and Statistics* 36: 175–182.

Jones, C.I. 2005. The shape of production functions and the direction of technical change. *Quarterly Journal of Economics* 120: 517–549.

Kaldor, N. 1961. Capital accumulation and economic growth. In *The theory of capital*, ed. F. Lutz. London: Macmillan.

Keynes, J.M. 1939. Relative movements of real wages and output. *Economic Journal* 49(193): 34–51.

Kravis, I.B. 1962. *The structure of income: some quantitative essays*. Philadelphia: University of Pennsylvania Press.

Kuznets, S. 1946. National income. In *Encyclopedia of the social sciences*, vol. 11. New York: Macmillan. Reproduced In *Readings in the theory of income distribution*, selected by a committee of the American Economic Association. Philadelphia: Blakiston.

Kuznets, S. 1959. Quantitative aspects of the economic growth of nations IV: distribution of national income by factor shares. *Economic Development and Cultural Change* 7(3,Part II): 1–100.

Ortega, D. and F. Rodríguez. 2006. Are capital shares higher in poor countries? Evidence from industrial surveys. Wesleyan economics working papers no. 2006-023, Wesleyan University.

Smith, A. 1776. *The wealth of nations: Books I–III*, 1986. Harmondsworth: Penguin Books.

Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39: 312–320.

Solow, R.M. 1958. A skeptical note on the constancy of relative shares. *American Economic Review* 48: 618–631.
Streightoff, F.H. 1912. *The distribution of incomes in the United States*, 1912. New York: Columbia University Press.

Labour-Managed Economies

B. Horvat

Labour management may be understood as a generic concept for all cases when enterprises are managed by those working in them. Institutional forms of such enterprises differ and also the degree of self-management varies. It may be expected that labour-managed firms and economies will behave differently from those run by capitalist or state managers.

The oldest labour-managed enterprises are producer cooperatives. Some of them survived from the Middle Ages; for example, monastic orders and some religious sects (e.g., Hutterites in the USA and Canada). The modern, non-religious equivalent are kibbutzim, which comprise about four per cent of the Israeli economy. They were preceded by various Owenite and Fourierist communities in the 19th century and coexist with a communitarian movement in Europe. Such cooperatives represent not only a specific organizational form but also a specific way of life, different from that of the rest of the community. Small communes in the developed countries and village communities in the non-capitalist environments also belong here.

The modern cooperative movement – cooperatives are just an organizational form of productive enterprises – was born in 19th-century Western Europe. At about the same time the first attempts were made to provide state capital to unemployed workers who were to run their enterprises by themselves (the Ateliers Nationaux of Louis Blanc in 1848).

Since the Paris Commune of 1870 every genuine social revolution has generated strong demands and massive implementation of workers management. That meant the right of workers to self-management regardless of the ownership of capital (for the history of workers' management see Horvat 1982, pp. 109–173). Most of these attempts did not survive the revolution itself.

After World War II there was a virtual explosion of various forms of labour management and for the first time an entire national economy (Yugoslavia) was subject to workers' management.

Institutional Forms

Proceeding from the less inclusive towards more inclusive forms, one may distinguish three pure models:

(1) *Partnership or partial cooperative*. Partners are the founders and the owners of the cooperative. They manage the firm on an equal right basis. They employ other individuals who do not have ownership and management rights. Law and medical firms in the West and frequently organized along such lines.
(2) *Full cooperative*. The firm is owned by all of its members and every member has one vote in management decisions.
(3) *Worker managed enterprise*. Capital is socially owned which means that it is accessible to every member of society on equal terms. All workers participate in management on the basis one man one vote. The organization is based on the distinction between the two types of authority: professional and political. All workers, or their representatives in the Workers' Council, decide on the policy issues. Given the policy thus established, professional coordinators and other experts make their professional decisions. The Workers' Council has, naturally, full access to external expertise. In this way the organization is supposed to combine maximum democracy with maximum efficiency. The participation of all workers means capturing all information that is available within a firm.

Models 1 and 2 are based on collective ownership. Model 3 implies social ownership.

Degrees of Participation

At around World War I the autocratic organization of typical capitalist firms began to encounter strong resistance. The need to expand war production and avoid strikes induced governments and employers of belligerent countries to experiment with some mild forms of workers' participation. Although similar attempts were made earlier, particularly in Germany, British *joint consultation*, as exemplified in the Whitley councils of 1917, may be taken as a landmark. Joint consultation means that the employer is obliged to consult his employees before making decisions that affect their work and income in some important way. However, the final decision is his.

The next step towards democratization of management was made in Germany after World War I when *codetermination* was introduced. Under the pressure of the 1918 revolution, when German workers demanded the socialization of the economy, the Weimar constitution envisaged codetermination. But this constitutional provision was never enacted. After the last war a series of laws were passed providing for workers' participation on the boards of directors – in some industries on a parity basis – and also reserving the post of the personnel director for the trade union representative. Today all West European countries, and many others as well, have some form of co-determination.

Further development led towards full-fledged *workers' management*. It was both revolutionary and reformist. As a result of a social revolution, workers' management was established in Yugoslavia (1950). The reformist way (called *democracia social de participatión plena*), was pioneered by Peru in the 1970s under President Velasco Alvardo, but the development was mostly reversed after his death. The same idea was taken over and more successfully implemented by the Swedes in the 1980s (Meidner 1978). Genuine democratization of management requires also a change in property relations; workers must have control over invested capital, at least partly. Swedish Wage Earners Funds are financed by a certain percentage of annual gross profits and payroll tax. They buy shares in the companies and are controlled by the unions. That, of course, is not full workers' management. The economy is still privately owned and unions are centralized organizations. But the Swedish reform marks a successful beginning of a reformist transition period.

Social Ownership

In the tradition of the First and Second Internationals, Soviet legal theory – and many authors elsewhere – identify state ownership with socialism. Thus the Soviet Civil Code of 1922 distinguishes three types of ownership, in ascending order: private, cooperative and state. The last one represents the basis for socialism. After a while it was discovered that the position of the worker in state firms is no different from that in private firms. Occasionally it may even be worse, since the state is a monopoly employer. Under both regimes the intra-firm hierarchy is preserved and management has autocratic power. Thus one has to distinguish the state ownership that characterizes the social order called *étatism*, from the social ownership which is appropriate for socialism, the latter being a full-fledged worker-managed economy.

Economic and legal theory of social ownership is still in its infancy and is virtually unknown outside Yugoslavia. The basic ingredients of the existing theory are as follows.

As a social category, ownership had three dimensions: *legal, social* and *economic*. In the *formal legal* sense social property is a bundle of rights intended to regulate economic transactions. Traditionally the inventory of such rights consisted of *ius utendi, fruendi et disponendi*. As a result of a long historical process, these rights came to be subject to four types of restrictions: (1) market restrictions – cartels are forbidden, monopolies will be broken up, prices are often regulated, etc.: (2) work restrictions – the length of the working day and week is regulated and certain safety measures are mandatory;

(3) ecological restrictions; (4) systemic restrictions – the value of productive capital cannot be reduced regardless of the sources of finance. Restrictions 1–3 are common for all modern societies, though they vary in comprehensiveness. Restriction 4 is specific for socialism.

The *social* dimension implies three rights: (1) every member of society has a right to work; (2) every member of the society has a right to compete for any work position if he meets the requirements of the work place; (3) every member of the society has the right of participation in management on equal terms.

Economically social ownership means that income from property (interest; land, mining, location and monopoly rents) belongs to society. Since income is the result of only three factors of production: natural resources, produced resources (capital) and labour – the first two are socially and the last one privately owned – the property right *usus fructus* implies income from live labour exclusively while everything else is capital income. The right (to the product of one's own labour) and the restriction (nothing except the product of labour) is the basis for the principle of distribution according to work. The right to income from capital implies that society is an economic owner of the entire social capital. The attribute *economic* means that formal legal ownership is largely irrelevant as long as the social and economic dimensions of social ownership are preserved. In other words, family farming and smallscale private (in the legal sense) production generally is fully consistent with socialism when it is worker managed and labour-managed economies generates income from work only (Bajt 1968). Income from work includes also income from entrepreneurship.

If we take into account that ownership relations determine particular social orders, then social ownership generates workers' management and distribution according to work (and vice versa) which are the basic constituents of socialism. An historical analysis of social revolutions shows that all of them have been motivated by the quest for justice, which has been interpreted as liberty, equality and solidarity. The three components of justice imply each other and we may take any one as a starting analytical concept. If we take equality as our guiding principle, a society will be considered egalitarian if its members are equal in their fundamental social roles. There are only three such roles: each of us is a producer, a consumer and a citizen. Equality of producers implies workers' management and social property; equality of consumers implies distribution according to work; equality of citizens implies a deconcentration of political power which is a pre-condition for political self-government.

We have arrived at a consistent social theory. Workers' management is a product of historical developments, ethical motivations and organizational solutions required for a society which is about to enter the 21st century. This is the conceptual frame within which we may now proceed to consider the micro and macroeconomics of labour management.

Microeconomics

A few years after the initiation of workers' management in Yugoslavia, an American graduate student, Benjamin Ward, selected it as the subject of his doctoral dissertation. He asked himself what could be the objective function of a worker-managed firm and wound up with the answer that it was not the maximization of profit but the maximization of income per worker (Ward 1958). This change in assumptions led to some very odd results. For a while Ward's paper passed unnoticed. Then the issue was taken up by Evsey Domar (1966), who considered the Soviet kolkhoz and introduced many inputs and a labour supply function into the analysis. Ward's misallocation effects were considerably weakened but not eliminated. The next step was an attempt at generalization in a book by Jaroslav Vanek (1970). He showed that free entry eliminates misallocation. However, since free entry is a long-run phenomenon, in the short run a labour managed firm will behave inefficiently. Vanek's book broke the silence of the profession. Soon there was a virtual explosion of papers and books and by now the bibliography has accumulated to many hundreds of items. A new discipline was born: the

economic theory of labour-managed firms. Yet, however sophisticated, the later contributions have not departed from the initial methodological framework. It has been taken as an established fact that a labour-managed firm (LMF) is less efficient than a capitalist-managed firm (CMF). Conservatives considered this as proof that capitalism was more efficient than socialism, while radicals tried to discover institutional conditions under which a LMF would catch up in efficiency with the CMF (e.g. reluctance to dismiss colleagues leads to a behavioural asymmetry and a different utility function). In the good neoclassical tradition only allocative efficiency has been discussed; the immensely more important productive efficiency has been hardly touched.

The essentials of the theory are as follows. A capitalist managed firm (CMF) maximizes *absolute* profit. Illyrian firm (IF) maximizes income *per worker*. For reasons to become apparent later I also add the worker managed firm (WMF) which maximizes income per worker *over a planning period*.

Consider a firm with a simple production function with two variable inputs, labour (x_1) and other resources (x_2),

$$q = f(x_1, x_2). \tag{1}$$

There is also fixed cost k, which may be interpreted as depreciation or as a capital tax. Profit appears as

$$\pi = pq - (wx_1 + p_2x_2 + k) \tag{2}$$

where p is the price of output, w is the wage rate and p_2 is the price of the other variable input. If profit is to be maximized, the first order conditions are the familiar marginal equations

$$\begin{aligned} \frac{\partial \pi}{\partial x_1} = 0, \quad &\rightarrow pq_1 = w \\ \frac{\partial \pi}{\partial x_2} = 0, \quad &\rightarrow pq_2 = p_2. \end{aligned} \tag{3}$$

The second order conditions are satisfied if diminishing returns are assumed, as will be done throughout.

An analysis of conditions (3) shows that: (a) an increase in product price increases output and employment; (b) an increase in factor prices decreases output and employment; (c) a change in fixed cost produces no effect, since k does not appear in the conditions; and (d) labour is treated the same as any other resources, there is complete symmetry.

Let us now replace capitalist management by a worker's council. Since wages do not exist, we cannot establish profit. As already mentioned, the objective function is now income per worker

$$y = \frac{pq - (p_2x_2 + k)}{x_1}. \tag{2a}$$

Ward was not quite sure that the actual Yugoslav firm maximized y, and so he preferred to talk about the 'Illyrian firm'. The first-order conditions are now

$$\begin{aligned} \frac{\partial y}{\partial x_1} = 0, \quad &\rightarrow pq_1 = \frac{pq - (p_2x_2 + k)}{x_1} = y \\ \frac{\partial y}{\partial x_2} = 0, \quad &\rightarrow pq_2 = p_2. \end{aligned} \tag{3a}$$

It is evident that the second-order conditions are also satisfied.

We cannot analyse (3a) directly. I shall therefore rearrange terms

$$q - q_1x_1 = \frac{k}{p} + \frac{p_2x_2}{p}. \tag{4}$$

It is easy to see that the following is true

$$\frac{\partial}{\partial x_1}(q - q_1x_1) = -q_{11}x_1 > 0. \tag{5}$$

A similar analysis now produces the following results: (a) an increase in p reduces the right-hand side of equation (4); in order to preserve equilibrium, the left-hand side must also be reduced, which according to (5) amounts to reducing employment x_1 and, consequently (by virtue of (1) above), output; (b) an increase in the factor price of other resources has the same effect as in

Labour-Managed Economies, Table 1 Effects of various changes on output and employment

Type of change	CMF	IF	WMF
Increase in product prices	+	—	+
Increases in wages	—	0	0
Increase in the price of material inputs	—	—	—
Increase in fixed cost	0	+	0

the neoclassical firm; (c) an increase in fixed cost k increases output and employment; and (d) factors are not treated symmetrically, since wages do not occur in (3a) and the conditions are structured differently.

The entire exercise is more clearly surveyed in Table 1.

By treating labour differently from material inputs, Illyrians behave in a strange way and impair the efficiency of their firms. When product prices in the market increase, they reduce output. The economy is thus hopelessly unstable. When the government wants to increase employment, it must levy a lump sum tax. The higher the tax, the higher the output and employment. Wage policy is of no use, since Illyrians disregard wages. Because $y > w$, and q_1 (Illyrian) $> q_1$ (capitalist), where q_1 is the marginal product of labour, an Illyrian firm employs fewer workers and produces less than its capitalist counterpart. For the same reason, it uses more capital than necessary. Less employment and higher capital intensity imply, for a given time preference, a smaller rate of growth.

Any meaningful theory must pass two fundamental tests: the verifiability of assumption test and the predictability test. A theory may pass both tests and still not be a correct one. If it fails to pass one or both of them, it is surely not satisfactory. If its assumptions cannot be verified, the theory has no explanatory power; if its predictions are wrong, it is simply useless. The latter test is much simpler and more conclusive, and so we may consider it first. For this purpose we rely on empirical research concerning the Yugoslav economy.

The theory predicts that an increase in price will reduce output. Nothing of the kind has been observed. Increases of price, as signals of unsatisfied demand, have been followed rather quickly by efforts to increase supply.

The theory also predicts that a reduction in k will reduce supply. When the 6 per cent capital tax was abolished in Yugoslavia in the 1960s, no one observed the predicted effect.

The theory predicts that the worker-managed economy will be labour saving. The Yugoslav experience shows, on the contrary, chronic overemployment in the firms.

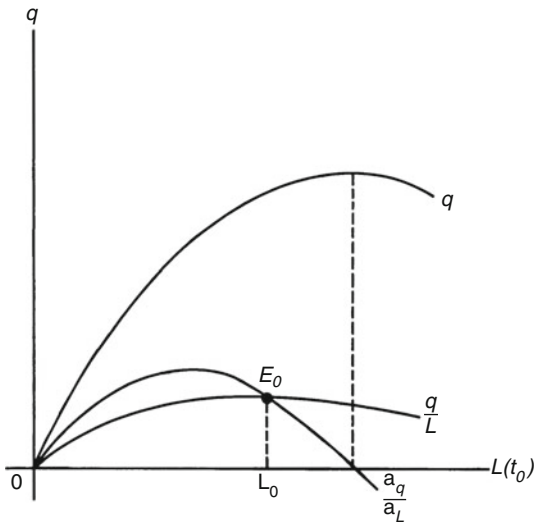
Where saving and investment are concerned, the theoretical prediction is again wrong. Internal saving of the firm is modest (which is explained by a negative interest rate), but borrowing is enormous, so that the national saving rate oscillates around 35 per cent of GNP (with government accounting for a negligible share). On the other hand, overinvestment tends to contribute to chronic inflation. Social property and planning reduce risks and so increase investment opportunities.

The formal reason for the supposedly perverse behaviour of the Illyrian firm is to be found in the form of the objective function, which is a ratio. If a CMF were assumed to maximize the *rate* of profit, it would display symmetrical perverse effects (Dubravčić 1970). Alfred Marshall avoided such consequences by distinguishing between the short and the long run; in the short run capital is assumed fixed and so maximizing profit and maximizing rate of profit comes to the same thing. Horvat (1969, 1985) suggested a similar device, which becomes available after a serious methodological error in the existing literature is eliminated. The error consists in deriving dynamic behavioural consequences from static assumptions.

If technology is fixed, we may assume that time does not matter. The resulting traditional static production function implies discovering output possibilities from varying quantities and proportions of inputs. If, however, we accept as a fact of life that technology is changing all the time, output will be a function of inputs *and* time

$$q = f(x_1, \ldots, x_n, t). \tag{6}$$

Marginal product in a production function thus defined is *not* a partial derivative of output



Labour-Managed Economies, Fig. 1 Neoclassical production function (technology fixed)

with respect to one of the inputs, $\partial q/\partial x_i \neq MP_i$. Thus, the routine maximization procedure is meaningless. Even treating t as a shift parameter will not do. The production function not only shifts in time but also *changes its shape*. Besides, if capacity is not fully used (i.e. less than 3 shifts), which is a normal situation, the returns to variable factors are as a rule increasing.

In Fig. 1 the law of variable proportions is operative and marginal product of labour is *diminishing*. With *fixed technology* we examine changes of q due to changes in L. E_0 represents maximum per worker output for technology known at t_0. Since technological progress is a *positive* function of time, marginal product of *dated* labour is *increasing*. Thus new workers, as well as the intramarginal ones, are more productive and per worker income is increasing. The invented perversities of the Illyrian firm disappear.

What the worker managed firms actually do consists in solving dynamic programmes of the following type: maximize total wage income of the currently employed workers over the agreed upon planning period of n years under a set of some six constraints (not all need be binding):

(1) All new workers will be given the same wage.
(2) Wage less than $\overline{w}_t - a_t$ ($\overline{w}_t$ is the average for the economy, a_t = collectively determined welfare factor) will not be tolerated.
(3) Wages higher than $\overline{w}_t + b_t$ are not desirable, because then the social pressure on the firm's funds becomes unbearable (local football club, local welfare programmes, etc.). Besides, progressive taxation drains too many resources away.
(4) Income distributed in wages is progressively taxed, income invested ('profit') is not. Society has no reason to tax its own capital; on the contrary.
(5) Bank investment loans are given under the condition that c per cent of investment finance is provided out of the firm's funds which serves as a collateral. Thus not all income accrues to wages, but part of it must be saved.
(6) Since capital represents social property, it can only be augmented and never eaten up whatever the source of finance. This solves the problem of the terminal stock of capital.

Once this programme has been solved, the aspiration wages (w^*) becomes known for the current decision period. The firm now maximizes the short run surplus

$$\max\ pq - w^*L - \left(\sum p_i x_i + k\right). \qquad (7)$$

At the end of the accounting period the actual wage is likely to be different, the difference $w - w^*$ depending on the business result. As the actual wage depends on the *ex post* results, it does not appear in the *ex ante* maximization conditions. Since (1) part of income is not distributed, (2) workers are not owners of capital but (3) capital investment is a precondition for increases in wages, it makes no sense to maximize *per worker* surplus or, which is the same thing, total per worker income. Equation (7) is mathematically identical to (2) and so neoclassical efficiency requirements are satisfied as is shown in Table 1.

One additional objection has been raised by Eirik Furubotn and Svetozar Pejović (1970). If workers invest in their firm, they benefit from

the increases in wages. If they put their money in the bank, they will collect not only interest but also principal at some future date. Unless the rate of profit is sufficiently higher than the bank rate of interest and the planning horizon sufficiently long, workers will distribute the entire income and investment will be reduced. The objection has some force in the case of cooperatives in a capitalist economy and for this reason the impressive Basque Mondragón cooperative system introduced personal capital accounts for its members. The accounts function similarly to bank deposit accounts. In a fully socialized economy no problem arises. Aggregate investment is a matter of social plans. Whether workers save directly as producers, or indirectly (via banks) as consumers, saving will be used to finance investment. However, worker managers have very strong incentives to save directly as producers because (a) such savings are free of tax while personal incomes and, consequently, savings from such incomes, are progressively taxed (which easily overcompensates the Furubotn–Pejović effect) and (b) the greater are firm's own funds, the greater is the independence in the decision making. Either bank control is avoided or larger bank loans become available and in both cases worker managers find it easier to expand production and insure their wages against competition.

Macroeconomics

Since empirically based macroeconomics is possible only when at least one national economy exists, macroeconomics of worker management is much less discussed and is almost entirely based on the Yugoslav institutions. Consequently, unlike in microeconomics, no well-developed theory – correct or fallacious – has appeared so far. In what follows some of the more important results will be presented.

Business Cycles

Even with perfect foresight, adjustments are not instantaneous. Mathematically formulated lagged adjustments lead to characteristic equations with real or complex roots depending on the parameters. Economic parameters seem to be such as to generate complex roots, that is, oscillations. However, even if parameters were to guarantee stability, external shocks (changing weather conditions, changing international environment, etc.) would initiate cycles, as Ragnar Frisch recognized long ago. The procedure may be reversed and, instead of modelling individual processes, an autoregressive scheme for the social product may be assumed right from the beginning and the relevant parameters estimated. For the Yugoslav economy, the parameters appeared significant for two cycle paths: the strongly damped short cycles (3 to 9 quarters) were superimposed over the longer regular ones (10 to 17 quarters) with the multiple correlation coefficient exceptionally high, $R = 0.93 - 0.98$ (Horvat 1969, pp. 215–20). This looks very much like Schumpeter's Kitchins and Juglars.

Compared with what is known about business cycles in the capitalist economy, Yugoslav cycles have some specific features. The accelerator is not operative; acceleration or retardation in *production* leads to breaks in investment activity, not the other way round. Inventories are depleted in the upswings and piled up in the downswings. This is explained by the reluctance of worker managers to dismiss their colleagues and the willingness of banks to finance inventory accumulation. The inverse movement of inventories has a significant stabilizing effect. Pressure on prices is less at high rates of growth and greater at low rates. Consequently prices fall or rise more slowly in times of expansion (positive excess demand) and there is inflationary pressure in recession periods (negative excess demand). Movements of credit either do not explain price changes or an increase in money supply reduces prices. This paradox is easily resolved when one remembers that credit stimulates production, expansion of production lowers costs, lower unit costs put less pressure on prices and so it appears statistically that credit lowers prices.

Planning and market, contrary to widespread beliefs, are not antithetical but complementary. Since the market is inherently unstable, planning

is indispensable for its normal functioning. On the other hand, in the implementation of economic policy, the market is the most efficient planning device. Within this conceptual framework planning means that strategic proportions (such as the volume, structure and regional allocation of investment) are realized, which is known as 'planning of global proportions'. The social plan, which includes also non-economic goals, is a kind of Rousseau's 'Social Compact'. It has four basic functions: The plan is above all a *forecasting instrument*; by generating information, it reduces uncertainty. As such it is an *instrument for the coordination of economic decisions*. The social plan is prepared in a participatory fashion, which implies prior harmonization of development goals of industries and regions. As such it provides the basis for the economic policy and so it serves an *instrument for guiding economic development*. As an elaboration of economic policy, the plan represents an *obligation for the body that has adopted it* and a directive for its organs. Other economic agents are free to make their decisions themselves which, of course, is a precondition for genuine workers' management and a free market.

Distribution According to Work

Distribution of income passes through two stages: first the firm earns income and then it distributes income among the workers. Workers themselves decide on the internal distribution of income (the structure of wages and the share of accumulation). Total income earned depends on their work and entrepreneurship, but also on general market conditions. It is the duty of the planning authorities to equalize starting business conditions for all firms. This may be done in the following way (Horvat 1982, pp. 263–282). All plants are classified into relatively homogeneous industry groups comprising twenty or more units. It may be assumed, for statistical reasons, that all industry groups are about equal in terms of effort and entrepreneurship and so average per worker income ought to be equal for all groups. If that is achieved, intragroup wage differences reflect exclusively distribution according to work. It remains to establish an objective standard for the measurement of average group incomes. Since capital is social, the planning authorities charge a uniform interest rate. Land, mining and locational rents are extracted in the usual way. An occupation, which is performed under approximately the same conditions throughout the economy, is taken as a standard unit. Incomes of all firms are expressed in such units using each firm's own wage differentials as weights. If wages thus aggregated differ from one industry group to the other, the planning authorities must adjust policy instruments in order to achieve the highest possible degree of equality. The remaining (extra) profits represent monopoly rents and are subject to progressive income taxation. Although industry averages are about equal – unless there is some reason for stimulating the development of a particular industry – differences between individual firms may be great and that provides incentives for work effort and entrepreneurship.

Finally, we may mention two classical problems – optimum distribution of income and optimum investment – which have proved analytically intractable under individualist or étatist institutions of privately or state owned economies. If firms and states are hierarchically structured and autocratically managed, there is in principle no possibility for interpersonal comparisons. If, however, all concerned participate in the decision making and a consensus is achieved, there is no possibility of improving upon such a solution. In a class structured society, consensus is in principle impossible. The higher the wages, the lower the profits and vice versa. In a classless society it is at least logically admissible.

See Also

- ▶ Codetermination and Profit-Sharing
- ▶ Command Economy
- ▶ Cooperatives
- ▶ Market Socialism
- ▶ Prices and Quantities

- ▶ Principal and Agent (i)
- ▶ Socialist Economies

Bibliography

Bajt, A. 1968. Social ownership – Collective and individual. Reprinted and translated in *Self-governing socialism: A reader*, vol. 2, ed. B. Horvat, M. Marković, and R. Supek. New York: Sharpe, 1975.

Domar, E.D. 1966. The Soviet collective farm. *American Economic Review* 56: 734–757.

Dubravčić, D. 1970. Labour as an enterpreneurial input: An essay in the theory of producer cooperative economy. *Economica* 37: 297–310.

Furubotn, E., and S. Pejović. 1970. Property rights and the behaviour of the firm in a socialist state: The example of Yugoslavia. *Zeitschrift für Nationalökonomie* 30: 431–454.

Horvat, B. 1969. *Business cycles in Yugoslavia*. Trans. H.M. Kramer. New York: M. Sharpe, 1971.

Horvat, B. 1982. *The political economy of socialism*. New York: M. Sharpe.

Horvat, B. 1985. The theory of the worker managed firm revisited. *Journal of Comparative Economics*. (The basic idea was first propounded in 'Prilog zasnivanju teorije jugoslavenskog poduzeća', *Ekonomska analiza*, 1967, 7–28.)

Meidner, R. 1978. *Employee investment funds*. London: Allen & Unwin.

Vanek, J. 1970. *The general theory of labour-managed market economies*. Ithaca/New York: Cornell University Press.

Ward, B. 1958. The firm in Illyria: Market syndicalism. *American Economic Review* 48(4): 566–589. In *The socialist economy*, ed. B. Ward. New York: Random House, 1967.

Labour-Managed Firms

Louis Putterman

Abstract

Labour-managed firms (LMFs) are enterprises over which suppliers of labour hold full control rights. Theoretical analysis suggests that such firms will behave in a distinctive and sometimes 'perverse' manner in response to short-run changes, but richer models can reverse the more problematic results, and the simple model indicates that LMFs behave no differently from capitalist firms in long-run competitive equilibrium. Empirical studies indicate that LMFs, while uncommon in most market economies, can achieve high productive efficiency. The search for an understanding of why LMFs are relatively rare has contributed to both positive and normative economic analysis.

Keywords

Codetermination; Democracy; Efficiency wages; Firm, theory of; Free-rider problem; Labour-managed firms; Mill, J. S.; Ownership and control; Partnerships; Profit sharing; Socialism; Total factor productivity; Wage differentials

JEL Classifications

J0

Although the traditional theory of the firm gave little attention to institutional detail, the common assumption about the units that engage in the production and sale of goods and services was that they are owned and controlled by individuals who provide risk-bearing capital and who hire the services of workers as one among several variable inputs. Worker-run cooperatives had existed in small numbers at least since the Industrial Revolution, but the study of such firms using formal analytical tools awaited the added stimuli provided by the challenge of understanding collective farm performance in the Soviet Union and China, and Yugoslavia's experiment with worker-managed market socialism. The models developed in the late 1950s and thereafter were subsequently applied not only to those cases but also to understanding worker-owned firms in industrial market economies, to investigating hypothetical economies consisting exclusively of worker-run firms, and to attempting to explain why worker control is relatively rare. As studies on the topic multiplied, the term 'labour-managed firm' (LMF) came to be used by economists to describe an enterprise that

operates under the ultimate control of those who work in it.

Such a definition of an LMF permits considerable variation in other dimensions. To qualify as an LMF, for example, an enterprise's workers must have control in the sense that managers are appointed and can be removed by them or by their representatives. But the degree of direct worker involvement in decision-making can vary, from the more direct democracy of small cooperatives to the representative structures of large Mondragon cooperatives or the now-defunct Yugoslav firms. A frequent assumption is that the exercise of worker control follows 'one worker one vote' lines, but the LMF concept has sometimes been extended to firms that include a class of workers lacking control rights. Most importantly, perhaps, the term LMF has been applied both to firms in socialist economies, in which the private ownership of capital is prohibited and the enterprise's capital is the property of 'society' or of a collective, and to worker-owned firms in capitalist economies, in which individual workers can hold property rights in their enterprise's assets, for example through 'partnership deeds', 'individual capital accounts' or shares.

The principal example of an LMF with 'social capital' was the Yugoslav social enterprise, which arose from the application of new laws and principles to that country's Soviet-style state enterprises. Collective property was the prevailing legal notion applied to the land and equipment of collective farms in the Soviet Union, China, and other Communist states, and has also accounted for a portion of the assets of some Western worker-run firms. The canonical example of 'partnership deeds' is provided by worker-owned plywood companies in the United States. The capital account model was adopted by the group of worker-owned enterprises centered in the town of Mondragon in the Basque province of Spain. More hybrid cases with only elements of worker control, such as (*a*) the partial employee ownership of many American companies, (*b*) legal, medical, and other professional partnerships, (*c*) co-determination in Western Europe, and (*d*) the widespread employee ownership resulting from privatization programmes in many transition economies, also continue to stimulate interest in the economic analysis of firms run by workers.

Although the economic analysis of worker-run firms was stimulated by the cases mentioned, interest in the concept appears to be explained by other factors as well. Normative dissatisfaction with the capitalist employment relationship, in which workers assume a subordinate role in the production process and lack claims on enterprise profits, can be found among leading economists ranging from John Stuart Mill and Leon Walras to James Meade and Jacques Drèze. In his *Principles of Political Economy,* Mill, who dominated English political-economy in the mid-19th century, wrote

> To work at the bidding and for the profit of another, without any interest in the work – the price of their labour being adjusted by hostile competition, one side demanding as much and the other paying as little as possible – is not, even when wages are high, a satisfactory state to human beings of educated intelligence, who have ceased to think themselves naturally inferior to those whom they serve. (Mill 1848, pp. 760–1, n. 1)

He predicted the extinction of the capitalist firm ('There can be little doubt . . . that the relation of masters and workpeople will be gradually superseded by partnership', pp. 763–4) and opined that the result 'would be the nearest approach to social justice, and the most beneficial ordering of industrial affairs for the universal good, which it is possible at present to foresee' (p. 792). Modern political theorists such as Carole Pateman (1970) and Robert Dahl (1985) have argued that selfgovernment of the workplace by workers is an implied requirement of the principle of control of government by the governed, and that it would help to deepen democracy in more traditional political spheres.

Another source of interest in LMFs is the fact that the theoretical analysis of such firms promises insights into why the large majority of firms in market economies are established and controlled by investors rather than workers (Dow 2003). Whether that fact is to be attributed to

social custom, to the exercise of economic power by the wealthy, to aversion to risk by the poor, or to other factors, seems important for judging policies such as the expansion of codetermination or the use of worker ownership in future privatizations. It also has an important part to play in the ethical evaluation of the economic system as a whole.

The first wave of models of worker-management abstracted from issues of ownership and financing by assuming a fixed charge for capital or land, presumed to be rented by the firm but fixed in quantity in the short run. By contrast, the number of worker-members was taken to be variable, and the firm's main decision problem was to select a level of this input. In the seminal model of Ward (1958) and in subsequent treatments by Domar (1966); Vanek (1970); Meade (1972) and others, the objective was taken to be maximizing revenue per worker net of capital, land, or other charges. The first and most frequently noted finding of such models was that, with the maximand being the (endogenous or firm-specific) net earnings of a variable input, output might not respond normally to changes in the product price. In particular, Ward showed that, if labour is the only variable input, workers share net revenue on an equal basis, and the firm's objective is to maximize the earnings of each worker employed (without concern for workers who might have to be expelled to achieve earnings maximization for those remaining), then an increase in the product price would reduce optimal employment and thus the firm's output level. An industry consisting entirely of worker-run firms would accordingly exhibit a downward- rather than upward-sloping short-run supply curve, so that output would go down, rather than up, in response to increased demand (on the assumption that a short-run equilibrium is even possible). Labour would be misallocated among firms in the short-run equilibrium of a labour-managed economy, since those with high marginal product of labour would have no incentive to accept workers from those with low marginal product. As an added oddity, the firm would seek more workers if the cost of its fixed factor or a lump sum tax rose, and it would reduce its membership if the opposite occurred.

Long-run outcomes are less peculiar. Abnormal returns would attract new capital investments by existing firms and entry of other firms into the industry, giving the long-run supply curve a more conventional shape. In the very long run, with both the number of firms and their utilization of all factors being variables, equilibrium behaviour of labour-managed and conventional firms would be identical (Drèze 1976). Even short-run perverse supply responses would be rendered unlikely by a variety of factors. For example, Domar (1966) showed that the tendency of hypothetical LMFs to take on additional workers, as output prices fell or as net revenue was reduced by higher charges for fixed factors, could be annulled by incorporating in the model the supply of labour facing a firm. Other factors tending to weaken or reverse the 'perverse output supply response' include (*a*) use of variable inputs additional to labour, (*b*) flexibility of working hours, (*c*) reallocation of labour between product lines in multi-product firms, (*d*) reluctance to vote for the expulsion of incumbent members, perhaps because the voters face similar probabilities of being selected for expulsion, and (*e*) tradable membership rights.

Empirical research failed to provide evidence for backward supply responses by LMFs. Chinese collective farms were found to increase their output in response to higher government-set prices. Yugoslav firms were sometimes argued to be reluctant to take on new workers, in line with Ward model predictions, but no evidence has been adduced that they had insufficient flexibility over work hours or an inability to allocate workers among tasks and product lines so as to respond positively to better market conditions for a given product. In what is probably the most rigorous study of the supply response of worker-owned firms, that on US plywood cooperatives by Craig and Pencavel (1992), the authors concluded that the firms' output was significantly less responsive to product price changes than that of conventionally owned competitors, but they rejected

backward bending supply at high levels of significance.

Property rights and investment incentives were another major concern of the LMF literature beginning in the late 1960s. In Yugoslavia, workers were empowered to elect councils which selected and had governing authority over their companies' managers, but the capital stock of the company was legally owned 'by society', with workers having rights to current revenue but obligations to maintain and ideally to add to that stock. Furubotn and Pejovich (1970) demonstrated theoretically that with this rights structure self-interested workers would privately value new investments in their company only in so far as they expected to remain employed there and have their pay enhanced by the resulting higher productivity. For capital goods having a useful life exceeding the expected employment horizon of a worker, the privately appropriable rate of return must be adjusted downward to take into account truncation of the future earnings stream from the standpoint of the worker. Furubotn and Pejovich argued that Yugoslavia avoided an otherwise predicted dearth of investment only because government and Communist authorities continued to have considerable leverage over managers, and because the government encouraged companies to finance their investments with low-cost loans from the state banks, although this had the effect of pumping money into the economy and thereby fuelling inflation (Pejovich 1969).

Most economists studying the issue agreed that firms with social ownership of capital would suffer from a horizon problem of the sort that Furubotn and Pejovich identified. More generally, Vanek (1977) argued that failure to consider the scarcity price of capital can lead to inappropriate choice of technology, a factor that he viewed as being of sufficient importance to explain the historical failure of experiments with workers' management. He noted, however, that this need not be a general feature of LMFs. The truncation of the revenue stream that is considered when evaluating investments is a result not of worker control but of assuming that workers are deprived of any and all rights to their investments' returns after separation from their firm. The problem could thus be ameliorated or eliminated entirely by several methods, for instance the calculation of a severance payment based on the capitalized value of each worker's past contributions to their company's capital stock. Another possibility is for the worker to sell his position as a partner or member of the firm in a market. In a perfectly functioning membership market, the estimated remaining productivity or marketable value of physical and other assets created during the incumbent worker's career with the firm would be incorporated in the sale price of the membership right. Sertel (1982); Dow (1986), and Fehr (1993) demonstrated the theoretical ability of a membership market to eliminate the inefficiencies of worker control in other dimensions as well. Pencavel (2001) and Dow (2003), however, point out the rarity of such markets and evidence of their imperfect functioning, suggesting this as another place to search for possible explanations of why LMFs are not more common.

A much-discussed dimension of worker control and ownership is that of work incentives. Vanek argued that, as a means of motivating workers to give their full energies to their jobs, sharing profits is likely to be far superior to paying a fixed wage, since the worker on fixed pay receives the contractual wage regardless of how intensively she works and regardless of how the firm fares. At a theoretical level, such a claim can be disputed. On the one hand, the short-run insulation of the worker from the effects of her varying quantity or quality of effort need not imply the total absence of a connection, since the wage can be adjusted over time, including by performance-contingent promotions. Efficiency wage models also demonstrate the potential to elicit effort through the threat of firing for sub-par performance. A company's very survival may depend on the effort it obtains from its workforce. On the other hand, if workers share equally or according to predetermined proportions in the same pool of profit, then the incentive provided by profit-sharing suffers from the profit's dilution among many workers, and the prediction of a static or

finitely repeated model of effort choice is that rational workers will choose to free ride.

Despite this inconclusiveness of theory, empirical studies have given Vanek's intuition about profit-sharing and motivation more support than refutation. Profitsharing has often appeared to boost work incentives, in part because it changes the dynamics of worker–worker interactions – each worker now being far more inclined to show disapproval at a co-worker's slackness. The prevalence of mutual monitoring in worker-run firms is associated with concrete cost-saving from using fewer hired supervisors. Craig and Pencavel (1995) found total factor productivity to be between 6 and 14 per cent higher in worker-owned than in conventional plywood firms. Weitzman and Kruse (1990) found a positive effect of profit-sharing on productivity in a meta-analysis of studies of both worker-owned and conventional firms linking pay to profit. A similar finding is recorded by Doucouliagos (1995) in a meta-analysis of studies focusing on the effect of worker participation in decisionmaking.

If worker-run firms don't actually suffer from dysfunctional responses to changes in their economic environments, if they aren't dissuaded from investing by horizon problems, and if they motivate work effort at least as effectively as do conventional firms, why aren't they as common as Mill predicted they would one day be? Among the answers that have been proposed is that control by investors is superior to control by workers because investors' representatives can reach decisions more easily, the idea being that investors share a uniform objective of maximizing the firm's market value, whereas workers have multiple interests (job security, pleasant working conditions, higher earnings) upon which each may place a different weight, thus defying easy consensus (Hansmann, 1990). Another answer, suggested by Kremer (1997), is that less productive workers tend to use the firm's internal decision process to obtain a flatter wage dispersion, which weakens incentives for the more productive workers to stay with the firm. Still another possibility, formalized by Ben-Ner (1984) and Miyazaki (1984) based on an earlier suggestion by Mikhail Tugan-Baranovsky (1921), is that successful LMFs have an incentive to replace retiring members with non-member hired workers, concentrating the profits in the hands of a smaller member group which, in the limit, collapses to contain only one member, a proprietor. Studies of the life cycle of cooperatives, from creation to dissolution, find few cases following precisely this scenario, but situations in which workers sell their firm to private owners and become their employees are reported, for example, in the U.S. plywood sector.

Possibly the most promising place to search for explanations is in the area of financing. Because inputs are committed before output value is certain, and because time passes between the utilization of input services and the realization of revenue from product sales, firms typically need the services of both risk-bearers and financiers. There is no technical reason why all input suppliers, including workers, could not share in providing these services by accepting payments in the future and by working for shares of an uncertain total revenue, rather than for fixed wages. What is observed, however, is consistent with the view that the supply of riskbearing and financing services follows comparative advantage: specialists with greater willingness to bear risk and/or ability to pay for inputs up front become the suppliers of equity and debt finance, while workers are paid within short intervals in amounts promised in advance and not contingent on the firm's results. The fact that workers typically have less wealth and thus both less ability to supply funds or to finance their consumption from savings, as well as less willingness to bear risk, is likely to play an important part in explaining this (Putterman 1993). The thinness of potential markets for worker partnership shares and thus the absence or imperfection of the partnership market may add to the burden that financing their own firm imposes on workers (Dow 2003).

Although workers do accumulate substantial assets in pension funds in the United States, risk aversion (and pension fund regulations) may deter

them from investing too much of it in their own company or in any other single project. In a world in which wealth was quite equally distributed and was held mainly by workers, workers as principal owners of their own firms might still remain rare because workers might prefer to hold diversified portfolios containing shares of many firms other than their own.

If control (by managers) and ownership (by shareholders) are in any case separated in modern corporations, why not worker control with (outside) shareholder ownership? The fact that the de-linking of ownership and control remains incomplete even in those firms where ownership is most diffuse (in other words, the fact that shareholders retain ultimate control rights in publicly traded corporations) suggests an answer. Presumably ownership and control are almost universally linked in a market economy because the owner, the return on whose investment is subject to so many uncertainties, is unwilling to cede control over key decisions affecting that return. Until worker desires for control of their enterprises are strong enough that they are willing to bear considerable financial risk, or until market outcomes are altered by government interventions facilitating the de-linking of control rights from financial risk-bearing, LMFs appear likely to remain the exception to the rule in market economies.

See Also

- ▶ Domar, Evsey David (1914–1997)
- ▶ Meade, James Edward (1907–1995)
- ▶ Mill, John Stuart (1806–1873)
- ▶ Socialism
- ▶ Worker Participation and Profit Sharing

Bibliography

Ben-Ner, A. 1984. On the stability of the cooperative type of organization. *Journal of Comparative Economics* 8: 247–260.

Bonin, J., and L. Putterman. 1987. *Economics of cooperation and the labor-managed economy.* London: Harwood.

Bonin, J., D. Jones, and L. Putterman. 1993. Theoretical and empirical studies of producer cooperatives: Will ever the twain meet? *Journal of Economic Literature* 31: 1290–1320.

Craig, B., and J. Pencavel. 1992. The behavior of worker cooperatives: The plywood companies of the Pacific Northwest. *American Economic Review* 82: 1083–1105.

Craig, B. and Pencavel, J. 1995. Participation and productivity: A comparison of worker cooperatives and conventional firms in the plywood industry. Brookings Papers on Economic Activity – Microeconomics 1995, 121–160.

Dahl, R. 1985. *A preface to economic democracy*. Berkeley: University of California Press.

Domar, E. 1966. The Soviet collective farm as a producers' cooperative. *American Economic Review* 56: 734–757.

Doucouliagos, C. 1995. Worker participation and productivity in labor-managed firms and participatory capitalist firms: A meta-analysis. *Industrial and Labor Relations Review* 49: 58–77.

Dow, G. 1986. Control rights, competitive markets, and the labor management debate. *Journal of Comparative Economics* 10: 48–61.

Dow, G. 2003. *Governing the firm: Workers' control in theory and practice*. New York: Cambridge University Press.

Drèze, J. 1976. Some theory of labor management and participation. *Econometrica* 44: 1125–1139.

Drèze, J. 1989. *Labor management, contracts and capital markets: A general equilibrium approach*. Oxford: Blackwell.

Fehr, E. 1993. The simple analytics of a membership market in a labor-managed economy. In *Markets and democracy: Participation, accountability, and efficiency*, ed. S. Bowles, H. Gintis, and B. Gustafsson. Cambridge: Cambridge University Press.

Furubotn, E., and S. Pejovich. 1970. Property rights and the behavior of the firm in a socialist state: The example of Yugoslavia. *Zeitschrift fur Nationalökonomie* 30: 431–454.

Hansmann, H. 1990. The viability of worker-ownership: An economic perspective on the political structure of the firm. In *The firm as a nexus of treaties*, ed. M. Aoki, B. Gustaffson, and O. Williamson. London: Sage.

Kremer, M. 1997. Why are worker cooperatives so rare? Working Paper No. 6118. Cambridge, MA: NBER.

Meade, J. 1972. The theory of labor-managed firms and of profit-sharing. *Economic Journal* 82: 401–428.

Meade, J. 1989. *Agathotopia: The economics of partnership*. Aberdeen: Aberdeen University Press.

Mill, J.S. 1848. *Principles of political economy*, 1936. London: Longmans, Green and Co.

Miyazaki, H. 1984. On success and dissolution of the labor-managed firm in the capitalist environment. *Journal of Political Economy* 92: 909–931.

Pateman, C. 1970. *Participation and democratic theory.* Cambridge: Cambridge University Press.

Pejovich, S. 1969. The firm, monetary policy and property rights in a planned economy. *Western Economic Journal* 7: 193–200.
Pencavel, J. 2001. *Worker participation: Lessons from the worker Co-ops of the pacific northwest*. New York: Russell Sage Foundation.
Putterman, L. 1993. Ownership and the nature of the firm. *Journal of Comparative Economics* 17: 243–263.
Sertel, M. 1982. *Workers and incentives*. Amsterdam: North-Holland.
Tugan-Baranovsky, M. 1921. *Sotsialnyia Osnovy Kooperatsii*. Berlin: Slowo Verlagsgesellschaft.
Vanek, J. 1970. *The general theory of labor-managed market economies*. Ithaca: Cornell University Press.
Vanek, J. 1977. The basic theory of financing of participatory firms. In *The labor-managed economy: Essays by Jaroslav Vanek*, ed. J. Vanek. Ithaca: Cornell University Press.
Ward, B. 1958. The firm in Illyria: Market syndicalism. *American Economic Review* 68: 566–589.
Weitzman, M., and D. Kruse. 1990. Profit-sharing and productivity. In *Paying for productivity: A look at the evidence*, ed. A. Blinder. Washington, DC: Brookings Institution.

Laffer Curve

Don Fullerton

Abstract

A Laffer curve is a hump-shaped curve showing tax revenue as a function of the tax rate. Revenue initially increases with the tax rate but then can decrease if taxpayers reduce market labour supply and investments, switch compensation into non-taxable forms, and engage in tax evasion. The revenue-maximizing tax rate can be calculated from an estimate of the elasticity of taxable income with respect to the after-tax share. Some studies find this elasticity to be near zero, and others find it to exceed 1. The mid-range for this elasticity is around 0.4, with a revenue peak around 70 per cent.

Keywords

Capital supply; Elasticity of labour supply; Elasticity of taxable income; Excess burden of taxation; Home production; Income effect; Labour supply; Laffer curve; Leisure; Marginal and average tax rates; Progressive and regressive taxation; Revenue maximization; Substitution effect; Supply side economics; Tax avoidance; Tax compliance; Tax evasion; Tax revenue; Taxation of corporate profits; Taxation of income

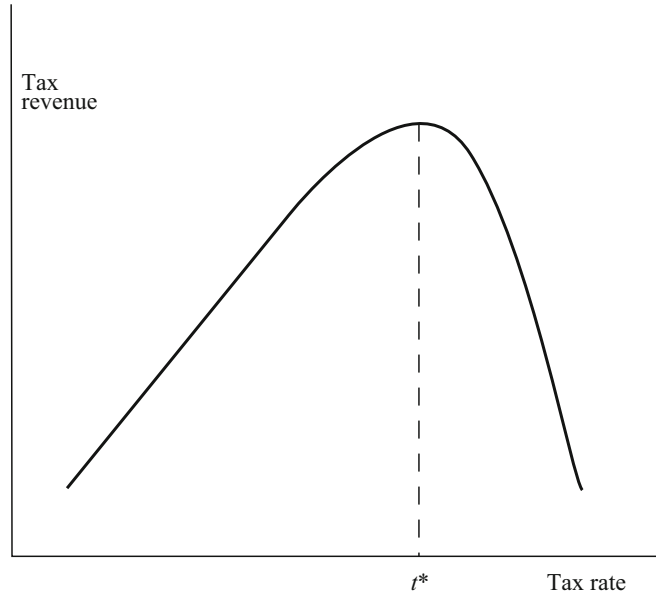
JEL Classifications

H2

On a napkin in a Washington restaurant in 1974, Arthur Laffer famously drew his hump-shaped curve showing tax revenue as a function of the tax rate (see Fig. 1). Revenue is zero both when the tax rate is zero and when the tax rate is 100 per cent or more. In between must be some t^* that maximizes revenue. The point is that taxes discourage supply of labour, especially by secondary workers in the family who have elastic behaviour, and they discourage supply of capital over time. Thus, proponents became known as 'supply siders'. So far, these points were well accepted, as economists are quite familiar with the idea of supply as well as demand. Even as far back as 1776, Adam Smith understood that 'High taxes, sometimes by diminishing the consumption of the taxed commodities, and sometimes by encouraging smuggling, frequently afford a smaller revenue to government than what might be drawn from more moderate taxes' (Smith 1776, V, II).

The more controversial claim was that the US tax rate was greater than t^*, on the 'prohibitive range' where no rational government would knowingly operate, meaning that a reduction in tax would actually increase government revenue.

Initial research focused on static models of labour supply. Stuart (1981) builds a simple analytical model with a taxed sector and an untaxed sector, and he chooses parameters to represent Sweden. The untaxed sector includes illicit tax evasion as well as leisure and home production such as painting your own home, growing your own vegetables, cooking your own meals, and cleaning your own house. He finds a peak at 70 per cent, which is fairly high, but then he also finds that Sweden has an overall effective



Laffer Curve, Fig. 1 The Laffer curve

marginal tax rate of 80 per cent! Then Fullerton (1982) describes two models. First, in a simple partial equilibrium model where η is the labour demand elasticity and ε is the labour supply elasticity, it is easy to show that $t^* = (\eta - \varepsilon)/[\ \eta(1 + \varepsilon)]$. If the labour demand curve is flat ($\eta = -\infty$), to focus on supply, then $t^* = 1/(1 + \varepsilon)$. Thus, higher ε implies lower t^*. The second model is a multi-sector computable general equilibrium model of the United States, but one that still requires an overall labour supply elasticity (ε). Based on estimates that are zero or negative for men and positive for women, the choice of $\varepsilon = 0.15$ in this model yields $t^* = 79$ per cent.

This research faces a number of problems. First, we do not really know the labour supply elasticity, and heterogeneity means we have no such thing as 'the' elasticity anyway. Second, we do not know the current tax rate either, since actual tax systems are complicated combinations of income, payroll, and sales taxes. For example, the payroll tax does not apply for workers whose tax payments are offset at the margin by additional expected social security benefits, and it also does not apply for those above the cap. Third, the income tax is progressive, which means different rates for different individuals. All this heterogeneity means no such thing as 'the' tax rate.

Fourth, even if we ignore heterogeneity, a progressive system means that the marginal tax rate (which affects incentives) exceeds the average tax rate (which affects revenue). Then the question of how a change in marginal tax rate affects revenue is not well defined, because one must also specify how the reform affects average rates. Even if an increase in all marginal rates raised revenue, for example, an increase in only the top marginal rate may not. Also, if a change in progressivity transfers money between groups, then the outcome depends on different *income* elasticities of labour supply. A reduction of the top marginal tax rate may seem to have the best potential for a Laffer effect if both (a) the rate is high and (b) those workers are elastic. But if part of the increased revenue comes from redistribution between taxpayers with different elasticities, then it is not a true Laffer effect.

Fifth, the Laffer curve itself is not well defined, with revenue on the vertical axis, because it matters how that revenue is spent. Interestingly, Malcomson (1986) shows that the Laffer curve may continue to slope upwards, all the way to a tax rate of 100 per cent, which would mean no prohibitive range at all! Yet Gahvari (1989) shows how this result depends on the assumption that revenue is used to provide a public good that is

separable in utility. Then the tax hike has an income effect that increases work effort, and revenue may continue to rise. If the increased revenue is used for lump-sum transfers, however, then this cash tends to offset the income effect, leaving only the substitution effect that is so emphasized by the supply siders in the first place.

So far, these models are static models of labour supply. Agell and Persson (2001) build a one-sector endogenous growth model with capital as the only input, and no labour at all, yet they obtain a strikingly similar result. They allow for separable government spending G or cash transfers T. One of their alternative definitions of a 'dynamic Laffer effect' is when government can reduce a tax rate and still increase at least one future year's G or T. They then show that a world with no transfers can never have a dynamic Laffer effect. The revenue-maximizing tax rate is 100 per cent, confiscating capital (so the growth rate is negative). With sizable transfers that are set to grow at some fixed rate, however, then a tax cut that increases the economy's growth rate means that transfers shrink as a fraction of GDP. Then, that negative wealth effect makes people save more, which increases the future tax base and may yield a dynamic Laffer effect.

The initial emphasis of the supply siders themselves was on supply of labour and capital, since these responses to a tax cut can increase income, growth, the tax base, and government revenue. Indeed, estimates of the labour supply elasticity mentioned above are estimates of the *hours* ' elasticity, the effect of the tax cut on hours worked. Yet what matters for tax revenue is the effect of the tax cut on 'taxable income'. Feldstein (1995) points out that a 'change in individuals' marginal income tax rates can induce them to alter their taxable income in a wide variety of ways, including changes in labour supply, in the form in which employee compensation is taken, in portfolio investments, in itemized deductions and other expenditures that reduce taxable income, and in taxpayer compliance' (1995, pp. 552–3). Thus begins a large empirical literature trying to estimate e, defined as the elasticity of taxable income with respect to a change in the marginal net-of-tax share $(1 - t)$. If the economy really had only a single tax rate t, then the revenue- maximizing tax rate is $t^* = 1/(1 + e)$.

Most of this literature takes a natural experiment approach that looks at years before and after a change in the income tax rate schedule, while comparing the top- bracket income group to the next-bracket income group. On the assumption that all other time trends affect the two groups similarly, then their e can be calculated by taking the difference between the two groups' change in reported taxable incomes compared with the difference between their changes in after-tax shares. Lindsey (1987) begins this literature by using cross-section data from the early 1980s for various income groups. The Economic Recovery Tax Act of 1981 reduced the top rate most, and the top bracket's reported taxable income increased the most. The implied elasticity is around 1.5, so the implied revenue-maximizing overall tax rate is around $t^* = 1/(1 + e) = 40$ per cent. This result stands in stark contrast to estimates mentioned above where t^* was 70–80 per cent.

This type of research also faces a number of problems. First, income inequality was trending upwards during these years, which might mean rising incomes at the top, relative to other groups, irrespective of the tax change. Second, random shocks to income mean that the top bracket may not contain the same individuals across years. Feldstein (1995) deals with this problem by use of panel data, tracking the same individuals before and after the top bracket rate cut of the Tax Reform Act of 1986. He also finds taxable income elasticities in excess of 1.0 (and sometimes 2.0 or 3.0).

Third, any given tax reform usually involves changes in the definition of taxable income, and not just changes in rates. Thus, these studies try to adjust their measure of income to use the same definition across years. Fourth, any change in the top personal income tax rate relative to the corporate tax rate might induce shifting: a change in personal taxable income that is offset by an opposite change in corporate taxable income. Fifth, the increase in taxable income in a single year after the tax change may be temporary rather than permanent. Sixth, the first few papers in this literature looked only at tax rate cuts in the 1980s, where other periods may have tax rate increases. Finally, each tax rate reform may involve a different set of

income tax rules that determine the ease of tax avoidance. In other words, there is no such thing as 'the' taxable income elasticity.

To deal with several of these problems, Goolsbee (1999) applies the natural experiment approach to *six* different tax reforms from 1920 to 1975, including both tax rate cuts and increases, and including periods with different trends in income inequality. He finds that the 1980s are atypical: 'the largest regression estimates of the taxable income elasticity from all of the previous historical periods are lower than the smallest estimates in the literature based on the 1980s' (1999, p. 43). Other studies find e around zero, as reviewed by Gruber and Saez (2002). They use a 1979–90 panel of tax returns to analyse *all* state and federal tax reforms during the 1980s, and they 'find that the overall elasticity of taxable income is 0.4, well below the original estimates of Feldstein but roughly at the mid-point of the subsequent literature' (2002, p. 3).

Finally, Kopczuk (2005) adds a measure of the tax base, relative to total income for each individual, and finds that it affects the estimate of the taxable income elasticity. In other words, that elasticity is not just a taxpayer's behavioural parameter, but depends on the tax code. The rich have a narrower tax base, and thus a higher elasticity. This also means that reforms to broaden the base can raise t itself (and reduce excess burden).

In summary, if you choose to oversimplify the world by using a single elasticity and a single tax rate, and if you ignore other problems above with the whole concept of the Laffer curve, then the recent mid-point estimate of $e = 0.4$ implies that tax revenue is maximized at $t^* = 1/(1 + e) = 71$ per cent.

See Also

- ▶ Labour Supply
- ▶ Tax Compliance and Tax Evasion
- ▶ Taxation of Income

Bibliography

Agell, J., and M. Persson. 2001. On the analytics of the dynamic Laffer curve. *Journal of Monetary Economics* 48: 397–414.

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.

Fullerton, D. 1982. On the possibility of an inverse relationship between tax rates and government revenues. *Journal of Public Economics* 19: 3–22.

Gahvari, F. 1989. The nature of government expenditures and the shape of the Laffer curve. *Journal of Public Economics* 40: 251–260.

Goolsbee, A. 1999. Evidence on the high-income Laffer curve from six decades of tax reform. *Brookings Papers on Economic Activity* 1999(2): 1–64.

Gruber, J., and E. Saez. 2002. The elasticity of taxable income: Evidence and implications. *Journal of Public Economics* 84: 1–32.

Kopczuk, W. 2005. Tax bases, tax rates and the elasticity of reported income. *Journal of Public Economics* 89: 2093–2119.

Lindsey, L. 1987. Individual taxpayer response to tax cuts, 1982–1984: With implications for the revenue maximizing tax rate. *Journal of Public Economics* 33: 173–206.

Malcomson, J.M. 1986. Some analytics of the Laffer curve. *Journal of Public Economics* 29: 263–279.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations.* London: J.M. Dent & Sons, 1975.

Stuart, C.E. 1981. Swedish tax rates, labor supply, and tax revenues. *Journal of Political Economy* 89: 1020–1038.

Laffont, Jean-Jacques (1947–2004)

Jean Tirole

Abstract

Jean-Jacques Laffont was one of the great economists of the last quarter of the 20th century, with an encyclopedic mind in a time of intense specialization. He won widespread respect and recognition for his breakthroughs in both theory (including public goods, contract theory, and the regulation of natural monopoly) and econometrics. In addition, he was energetically engaged in institution-building not only in Europe but also in Africa, Asia and Latin America.

Keywords

Access prices; Adverse selection; Asymmetric information; Auctions; Bayesian implementation; Budget balance; Coase Theorem; Collective choice; Collusion; Contract theory;

Disequilibrium macroeconomic models; Econometric Society; Econometrics; Engineering cost models; European Economic Association; Industrial organization; Instrumental variables; Laffont, J.-J.; Less developed countries; Nonlinear simultaneous equations; Occupational choice; Partial equilibrium; Productivity; Public goods; Regulation of natural monopolies; Regulatory capture; Revealed preferences; Risk aversion; Theory of entrepreneurs

JEL Classifications
B31

Jean-Jacques Laffont was born in 1947 and died in 2004 in Toulouse. He was one of the great economists of the last quarter of the 20th century. He made breakthroughs in many fields within both theory and econometrics, which made him perhaps the last encyclopedic mind in the economics profession at a time when the rapid growth of knowledge pushes most researchers into intense specialization. His creative and prolific contributions brought him widespread respect and recognition, from presidencies of learned societies (Econometric Society, European Economic Association) to numerous prizes (including the Yrö-Jahnsson prize), honorary memberships in foreign learned societies, honorary degrees from several universities and invitations to give numerous prestigious lectures. Besides his academic contributions – the topic of this contribution – Jean-Jacques Laffont will also be long remembered for his selfless contributions to institution building in Europe and in particular Toulouse, where his warmth, devotion and energy allowed him, starting nearly from scratch, to create an enthusiastic and congenial research environment. In Africa, Asia and Latin America also, he encouraged young economists to work with him on frontier economics and helped build research centres.

Public Goods

After completing his Ph.D. at Harvard University in 1974, Jean-Jacques Laffont embarked on a celebrated research agenda on public goods, in collaboration with Jerry Green (culminating in their 1979 book) and later with Eric Maskin. A collective decision-making problem with n economic agents ($i = 1, \ldots, n$) who have quasi-linear preferences of the form:

$$u_i = v_i(a, \theta_i) + t_i$$

consists in selecting a policy a and transfers t_i for each configuration of taste parameters $\theta = (\theta_1, \ldots, \theta_n)$. An efficient policy $a^*(\theta)$ solves

$$\max_{\{a\}} \sum_{i=1}^{n} v_i(a, \ \theta_i).$$

A central issue is how to implement this efficient action through appropriate transfers when agents privately know their own taste parameters. Clarke (1971), Groves (1973) and Vickrey (1961) (CGV) had defined 'mechanisms', in which agents announce 'types' $\hat{\theta}_i$, the collective decision is $a^*\left(\hat{\theta}\right)$ and agent i receives a transfer of the form

$$t_i\left(\hat{\theta}\right) = \sum_{j \neq 1} v_j\left(a^*\left(\hat{\theta}\right), \ \hat{\theta}_j\right).$$

They had shown that such schemes would induce each agent to truthfully reveal her preferences $\hat{\theta}_i = \theta_i$, as she internalizes the consequences of her choices on the welfare of others. Green and Laffont (1977) showed that these mechanisms were, up to the addition of a function $t_i^o\left(\hat{\theta}_{-i}\right)$ which is independent of the announcement of the others, the only schemes in which truthful revelation is a dominant strategy. Laffont and Maskin (1980), pioneering the 'differentiable approach' to mechanism design, then showed that the transfers t_i^o were but constants of integration when the v_i are differentiable in a and θ_i.

A consequence of Green and Laffont's characterization was that dominant strategy public good schemes are inconsistent with budget balance $\left(\sum_i t_i = 0\right)$. This negative result shifted the profession's attention to the weaker requirement of Bayesian implementation, in which truth telling

is an agent's best response to the other agents' truth telling. Laffont and Maskin's (1979) pioneering work showed that inefficiency necessarily resulted from the stricter requirement that the budget be balanced for each configuration of preferences; their paper led the way to the equally pioneering paper of Myerson and Satterthwaite (1983) stating the generic inefficiency of bargaining processes under asymmetric information. These two papers thereby identified one important limitation of the Coase theorem.

Contract Theory

More generally, during the decade following his Ph.D. Laffont was involved in many of the developments of contract theory, from adverse selection to moral hazard, from single-agent partial-equilibrium to general equilibrium settings. Examples of this work include the definitive treatment of adverse selection with Guesnerie (1984), the first model of occupational choice in which Kihlstrom and Laffont (1979) built a theory of entrepreneurs based on heterogeneity in risk aversion, and the prescient piece with Green (1986) on limited scopes for misreporting (the report $\hat{\theta}_i$ is restricted to belong to a subset of types that depends on the true θ_i), in which they showed how to amend the revelation principle and derived some implications for the magnitude of distortions brought about by private information.

Regulation

A common application of incentive theory is to the regulation of natural monopolies. The first experiments with price caps in the mid-1980s and later with deregulation raised questions about what could be expected from such reforms and about their potential pitfalls. Starting with the 1986 paper on the power of incentive schemes and up to their 1993 book, Laffont and Tirole focused on these issues, modelling the objective of the regulated firm as (variants of)

$$u = t - C(\theta,\ e,\ q) - \psi(e),$$

where t is the firm's budget, C its monetary cost, $\psi(e)$ an increasing and convex non-monetary function of the effort e, θ a technology parameter unknown to the regulator and q the vector of outputs. While costs and outputs are observable, the firm can transform naturally low costs into shirking (or private benefits). For any abstract regulatory mechanism $\left\{q\left(\hat{\theta}\right),\ t\left(\hat{\theta}\right)\right\}$, expressing, as a function of productivity, the effort needed to reach a given cost level for given outputs and applying the envelope theorem, the regulated firm's rent's sensitivity to the productivity parameter is given by

$$\frac{du}{d\theta} = \psi'(e)\left|\frac{\partial e}{\partial \theta}\right|.$$

where $\delta e/\delta\theta$ measures the firm's ability to transform productivity gains into private benefits (for example, for a single output q and $C(\theta,\ e,\ q) = (C_0 - \theta - e)q,\ \ |\delta e/\delta\theta| = 1$). This condition provides the intuition for the incentive-rent extraction trade-off: high-powered incentives schemes – that is, schemes for which the firm bears a high share of its cost (inducing a high effort and therefore a high $\psi'(e)$) – necessarily leave large rents (large $u(\theta)$ s) on the table (this is the reason why price caps are often subject to political pressure for renegotiation). The 1986 paper provided sufficient conditions for a menu of linear contracts to be optimal.

Subsequent work focused on how the power of the incentive scheme is affected by concerns for quality, auctioning of incentive contracts, dynamics (the ratchet effect), and regulatory capture. Laffont and Tirole argued that a key enabler of political capture of the regulatory process is the asymmetry of information with the political principal (perhaps Congress, and certainly the citizens), and that the regulatory response to the threat of capture was low-powered incentives, as these reduce rents and therefore make the concerted manipulation of information by the firm and its regulator less attractive to them.

Later, Laffont and Tirole derived theoretical principles for the design of access prices, a key ingredient of the liberalization policy, in the case of one-way access to a bottleneck such as a local

loop, an electricity grid or a railroad network (1994) and, in collaboration with Rey (1998a; 1998b), two-way access, that is, access to mutual termination bottlenecks present in telecommunications or the internet.

Jean-Jacques Laffont was adamant about the ability of economic theory to help guide economic development, provided that the theory is properly adapted to reflect the specificities of the developing world. In his posthumous (2005) book, he did just that in the context of regulation. Characterizing less developed countries as countries with easy side transfers within families, ethnic groups and social networks, a lack of a constitutional control of government, a weak rule of law, a high cost of public funds, politically dependent regulators, and weak accounting structures, he systematically drew the implications for the design of regulation, from the power of incentive schemes to universal service obligations and a positive theory of privatization.

More Contract Theory

Convinced that collusion was a key determinant of economic outcomes and institutions, Jean-Jacques Laffont engaged in a thoughtful and seminal line of research on the methodology and implications of models of collusion, in particular in collaboration with David Martimort. Their 1997 paper developed a general approach for the analysis of collusion among n agents against a principal; an upper bound on the potential damage of collusive activities is obtained by introducing a fictitious coordinator (or cartel ringmaster in an auction) who (a) privately elicits the n agents' types $\theta_1, \ldots, \theta_n$), (b) dictates the agents' behaviours in the game designed by the principal, and (c) breaks even. This 'side mechanism' must be incentive compatible as well as individually rational (the agents must be willing to collude).

In their 2000 paper, Laffont and Martimort point at the dual impact of the 'commonness' of information among agents. A fundamental insight due to Maskin (1999) is that information held by multiple agents can often be elicited at very low cost by having economic agents compete, challenge each other's reports, exercise options, and so on. Maskin's insight has wide-ranging consequences for the use of the informational content of financial and labour markets, auctions, options and other commonly used elicitation mechanisms for the design of contracts and organizations. Laffont and Martimort argue that Maskin's insight is most potent when the schemes have integrity, that is, they are not vulnerable to collusion among agents; for it is precisely when agents have the same information that it is easy for them to collude. Put differently, informational asymmetries among agents hinder collusion. Faure-Grimaud, Laffont and Martimort (2003) show that delegation is an optimal response to collusion.

On the more applied aspects of side-contracting, Laffont and Martimort (1999) showed that the separation of regulators may make capture more difficult. Laffont and Meleu (1997) provided one of the first endogenizations of side transfers, and showed that reciprocal supervision provides an undesirable conduit for collusion.

Econometrics

Quite remarkably, Laffont also made key contributions to theoretical and applied econometrics. As a Harvard student, he collaborated with Jorgenson to produce one of the first methods for estimating nonlinear simultaneous equations, in particular extending and studying the efficiency of minimum distance and instrumental variable estimators, paving the way for Hansen and Hansen and Singleton's 1982 pioneering contributions. Gouriéroux, Laffont and Monfort (1980) is another important illustration of Laffont's contributions to nonlinear econometrics, this time motivated by the identification of simultaneous equation models with latent variables, and in particular disequilibrium macroeconomic models.

Later, Laffont was one of the pioneers of the new empirical industrial economics. He firmly believed in the importance of theory for imposing structural constraints in econometric estimation, and in the continuous back-and-forth interaction

between industrial organization theory and empirics. His first research along these lines (with Gasmi and Vuong 1992) is on the study of tacit collusion in price and advertising in the Coca–Pepsi duopoly. He then found in auctions and their clear extensive form a most favorable ground for structural econometrics in IO. Positing Bayesian equilibrium strategies and adding parametric restrictions allows the researcher to identify the underlying distribution of types and thus the structure of the model. For example, Laffont, Ossard and Vuong (1995) develop a simulated nonlinear least-squares method to estimate auctions with independent private values for a range of first- and second-bid mechanisms and apply it to eggplant auctions in the south-west of France.

Last, Jean-Jacques Laffont's was also interested in the engineering cost models (with Gasmi et al. 2002) as he viewed these as enabling a better regulation of, say, universal service obligations or access prices.

Selected Works

1974. (With D. Jorgenson.) Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Annals of Social and Economic Measurement* 3: 615–640.

1977. (With J. Green.) Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica* 45: 427–438.

1979. (With J. Green.) *Incentives in public decision-making*. Amsterdam: North-Holland.

1979. (With R. Kihlstrom.) A general equilibrium entrepreneurial theory of the firm based on risk aversion. *Journal of Political Economy* 87: 719–748.

1979. (With E. Maskin.) A differentiable approach to expected utility maximizing mechanisms. In *Aggregation and revelation of preferences*, ed. J.-J. Laffont. Amsterdam: North-Holland.

1980. (With C. Gouriéroux and A. Monfort.) Disequilibrium econometrics in simultaneous equations systems. *Econometrica* 48: 75–96.

1980. (With E. Maskin.) A differentiable approach to dominant strategy mechanisms. *Econometrica* 48: 1507–20.

1984. (With R. Guesnerie.) A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics* 25: 329–69.

1986. (With J. Green.) Partially verifiable information and mechanism design. *Review of Economic Studies* 53: 447–56.

1986. (With J. Tirole.) Using cost observation to regulate firms. *Journal of Political Economy* 94: 614–41.

1992. (With F. Gasmi and Q. Vuong.) Econometric analysis of collusive behavior in a soft drink market. *Journal of Economics and Management Strategy* 1: 277–311.

1993. (With J. Tirole.) *A Theory of incentives in procurement and regulation*. Cambridge, MA: MIT Press.

1994. (With J. Tirole.) Access pricing and competition. *European Economic Review* 38: 1673–710.

1995. (With H. Ossard and Q. Vuong.) Econometrics of first-price auction. *Econometrica* 63: 953–80.

1996. (With Q. Vuong.) Structural analysis of auction data. *American Economic Review* 86: 414–20.

1997. (With D. Martimort). Collusion under asymmetric information. *Econometrica* 65: 875–911.

1997. (With M. Meleu.) Reciprocal supervision, collusion and organizational design. *Scandinavian Journal of Economics* 99: 519–40.

1998a. (With P. Rey and J. Tirole.) Network competition: I. Overview and nondiscriminatory pricing. *RAND Journal of Economics* 29: 1–37.

1998b. (With P. Rey and J. Tirole.) Network competition: II. Price discrimination. *RAND Journal of Economics* 29: 38–56.

1999. (With D. Martimort.) Separation of regulators against collusive behavior. *RAND Journal of Economics* 30: 232–62.

1999. (With J. Tirole.) *Competition in telecommunications*. Cambridge, MA: MIT Press.

2000. *Incentives and political economy*. Oxford: Oxford University Press.

2000. (With D. Martimort.) Mechanism design with collusion and correlation. *Econometrica* 68: 309–42.
2002. (With F. Gasmi, M. Kennet and W. Sharkey.) *Cost proxy models and telecommunications policy: A new empirical approach to regulation*. Cambridge, MA: MIT Press.
2002. (With D. Martimort.) *The theory of incentives: The principal–agent model*. Princeton: Princeton University Press.
2003. (With A. Faure-Grimaud and D. Martimort.) Collusion, delegation and supervision with soft information. *Review of Economic Studies* 70: 253–80.
2005. *Regulation and development*. Cambridge: Cambridge University Press.

See Also

- ▶ Agent-Based Models
- ▶ Auctions (Empirics)
- ▶ Cartels
- ▶ Contract Theory
- ▶ Development Economics
- ▶ Public Goods

Acknowledgment I am grateful to Jacques Crémer, Marc Ivaldi, David Martimort and Eric Maskin for helpful comments

Bibliography

Clarke, E. 1971. Multipart pricing of public goods. *Public Choice* 2: 19–33.
Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.
Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1048.
Hansen, L., and K. Singleton. 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50: 1269–1286.
Maskin, E. 1999. Nash equilibrium and welfare optimality. *Review of Economic Studies* 66: 23–28.
Maskin, E. 2004. Jean-Jacques Laffont: A look back. *Journal of the European Economic Association* 2: 913–923.
Myerson, R., and M. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 28: 265–281.
Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16: 8–37.

Lagrange Multipliers

S. N. Afriat

Abstract

Lagrange's 'method of undetermined multipliers' applies to a function of several variables subject to constraints, for which a maximum is required. Lagrange's procedure avoids the arbitrary distinction between independent and dependent variables. The method involves further variables, the 'multipliers' associated with the constraints, which have importance in application to economic problems. Beside the value obtainable from a given resource, one might also wish to know the 'marginal value' obtainable when a unit of it is added. The Lagrangian method is therefore a natural tool of the 'marginalist revolution', and the multiplier concept underlies 'shadow price', 'implicit value' and similar expressions.

Keywords

Chain rule; Convex programming; Implicit function theorem; Kuhn–Tucker conditions; Lagange multipliers; Lagrangian function; Marginal revolution; Separating hyperplane theorem

JEL Classifications

C0

Lagrange's 'method of undetermined multipliers' applies to a function f of several variables x subject to constraints, for which a maximum is required. The constraints can be stated as $g(x) = q$ where the vector q is constant. Ordinarily one might distinguish independent and dependent variables under the constraints, and then by substitution for the dependent variables in f one has a function of independent variables whose derivatives must vanish. Instead Lagrange offered a procedure elegantly without the arbitrary

distinction between variables and more suitable for some applications. The idea of it has other ramifications, such as for analytical mechanics, calculus of variations and control theory, beside the economic optimization dealt with here. The method involves introduction of further variables u, the 'multipliers' associated with the constraints. With n function variables and m constraints we then have $m + n$ variables (x, u). Lagrange's method depends on $m + n$ relations he obtained to determine these, and so the n function variables x which are among them and should give the required maximum. The remaining m variables u, the 'undetermined multipliers', really are just as well determined. But originally they were just part of this device for determining a maximum and their values had no interest even if they could be determined.

The multipliers in fact have a further significance, as derivatives that tell how the maximum value varies as the constraints have variation from a variation of q. They therefore have importance in application to economic problems. For, beside the value obtainable from given resources, one might also wish to know the 'marginal value' of any resource, the extra value obtainable when a unit of it is added. The Lagrangian method is therefore a natural tool of the 'marginalist revolution' and the multiplier has become a part of economic language; it is also the concept that underlies 'shadow price', 'implicit value' and similar expressions.

The most typical economic maximum problem is formulated differently from that dealt with by Lagrange. Rather, the constraints have the form of inequalities, expressing that some resource availability must not be exceeded; also, functions involved have convexity properties required by diminishing marginal returns. The theory of such problems is different and does not depend on what we have for Lagrange's classical problem. Yet despite the essential difference there is an impressive similarity, from the role of 'multipliers', so one can think that here again is Lagrange's method in another shape. But about these multipliers in the new context quite new things can be said. In either case, classical or new, the required maximum is associated with multipliers enabling certain conditions to be satisfied. Here is similarity, but premises and conclusions related to such conditions in each case are different.

Though form brings the two lines together it is altogether a mistake to see coincidence, and rather it is proper to make the treatments entirely separate, instead of trying to deduce one from the other. The difference is well appreciated from the complete difference in proofs of main points. One requires the implicit function theorem, at least in a certain approach, or more simply just the chain rule, as here. The other, convex programming, requires instead the theorem of the separating hyperplane. Again, one is entirely concerned with differentiable functions while the other in its main part is not, though the differentiable case treated by H.W. Kuhn and A.W. Tucker is very familiar. Reassuring for the connection, there are special problems where both lines are applicable, and then the multipliers involved are identical. But even then more can be said about the multipliers than would come simply from the classical case. Our review of the classical and new multiplier theories will make clear the cleavages and connections. We will also see peculiar, and remarkable, features of the matter in the special context of linear programming. Following the ordinary method of distinguishing independent variables and eliminating dependent variables we can, without any other thought about it, arrive at Lagrange's method from consideration of the derivatives the multipliers happen to represent. In that way, beside other possible merit, the multipliers become at the same time identified with those derivatives. Though this is not a usual procedure, it is a counterpart for classical multipliers of an argument that is essential for the new multipliers of optimal programming theory.

It is convenient now to denote the n function variables by z, reserving x for independent variables among these. Lagrange's problem is to determine a maximum of $f(z)$ subject to m constraints, stated

$$g(z) = q. \tag{1}$$

Variables are column vectors, and all functions are understood to be differentiable, so for instance

g has an $m \times n$-derivative matrix denoted g_z with elements $g_{ij} = \partial g_i/\partial z_j$.

As necessary for z to be a maximum (or minimum, in any case a stationary point) Lagrange concluded that

$$f_z = ug_z \text{ for some } u, \tag{2}$$

in other words the n conditions

$$f_i = \sum_i u_i g_{ij} (j = 1, \ldots, n).$$

Together with the m conditions

$$g_i = q_i (i = 1, \ldots, m)$$

provided by the constraints (1) we have $m + n$ *Lagrange conditions* on the $m + n$ variables

$$u_i (i = 1, \ldots, m), \; z_j (j = 1, \ldots, n).$$

Lagrange's method depends on the idea that these $m + n$ conditions can be solved to determine the $m + n$ variables, and so the n variables z_j which are among these. Put in another way, the multipliers u_i can be eliminated (and so left 'undetermined') and the conditions obtained then solved for the z_j.

With independent and dependent variables x and y under the constraints, the variables have a partition $z = (x,y)$, and we have a function $f(x, y)$ under constraints $g(x,y) = q$ that determine y as a function $y = Y(x,q)$. Then $g[x,Y(x,q)] = q$ is an identity and so, by differentiation with respect to x,

$$g_x + g_y Y_x = 0, \tag{3}$$

and with respect to q, $g_y \, Y_q = 1$, and since from here g_y and Y_q are inverse matrices we also have

$$Y_q g_y = 1. \tag{4}$$

For any q the constrained values of f are described by $f[x, Y(x,q)]$ as x varies without restriction. The x-derivatives must vanish for a stationary point, that is

$$f_x + f_y Y_x = 0. \tag{5}$$

On the assumption that this condition determines a unique point x for any q, the stationary points for various q are described by a function $x = X(q)$. Then the corresponding stationary values of f are given by the function

$$F(q) = f[X(q), Y(X(q), q)],$$

with derivatives

$$\begin{aligned} F_q &= f_x X_q + f_y \left(Y_x X_q + Y_q\right) \\ &= \left(f_x + f_y Y_x\right) X_q + f_y Y_q = f_y Y_q \text{ by (5)}. \end{aligned}$$

Hence

$$\begin{aligned} F_q g_x &= \left(f_y Y_q\right) g_x \\ &= \left(f_y Y_q\right)\left(-g_y Y_x\right) \qquad \text{by (3)} \\ &= -f_y \left(Y_q g_y\right) Y_x = -f_y Y_x \qquad \text{by (4)} \\ &= f_x \qquad \text{by (5)}, \end{aligned}$$

and also

$$F_q g_y = \left(f_y Y_q\right) g_y = f_y \left(Y_q g_y\right) = f_y \quad \text{by (4)}.$$

It has now been seen that

$$F_q g_x = f_x, F_q g_y = f_y,$$

that is, $f_z = F_q g_z$, which is (1) with $u = F_q$. Thus we have Lagrange's conditions, together with the identification $u = F_q$ for the multipliers.

For any x, the existence of u so that (x, u) satisfy Lagrange's conditions (1) and (2) is the condition for x to be a stationary point. It is therefore necessary for x to be a maximum, or a minimum, and on its own not sufficient for x to be either. Solutions of Lagrange's conditions, if there are any, therefore provide all stationary points, possibly many, without information that any should be a maximum. However, should a maximum be known to exist and the conditions be found to have a unique solution (x, u) then x is known to be that maximum. This is a common circumstance with many applications and where the method has strength.

Given any stationary point x, such as could be found from a solution of Lagrange's conditions, and so obtained by a condition on first derivatives at x, one can possibly find out if it is a local maximum, or a maximum in some neighbourhood of x under the constraints, by an examination of further conditions bringing in higher derivatives of x. However, no conditions on derivatives simply at the point x will tell anything about x except in the local sense. There is no way of telling x is a global maximum simply from a satisfaction of some condition on derivatives at x, of any order. Of course in economics a maximum is significant only in the global sense. Fortunately, typical functions of economics have convexity properties that enable one to go further on the basis of local conditions. Connected with this, any stationary point of a convex, or concave, function is necessarily a global minimum, or maximum, so in such cases first order conditions are enough. This matter has a part in further theory of Lagrange multipliers in the more typically economic context of convex programming.

Lagrange's method can be described with reference to the 'Lagrangian function'

$$L(x,u) = f(x) - u[g(x) - g],$$

as requiring the x and u derivatives to be set to 0. This way of putting it is without significance except as a cook-book statement. One first learns about setting derivatives to zero when there are no constraints, and now even though there are constraints one can with confident familiarity do it again, even with the impression that the Lagrangian function should be at a maximum as if the recipe had that sense. There is better occasion for something like this in convex programming, where u is fixed so as to make x a maximum. A problem with inequality constraints is stated

$$(M)\ \ \text{Max}\ f(x) : g(x) \leq q,$$

functions being defined in a set A. It can be imagined that A is an *activity* set, and the performance of any $x \in A$ gives a return $f(x)$ and has a cost in terms of various resources given by the vector $g(x)$, so for *feasibility* this must not exceed the available stock q, so $g(x) \leq q$ is required. The problem is to find an *optimal solution,* an activity x that gives the greatest return attainable with the available resources, as asserted by the condition

$$M(x) =: g(x) \leq q\,, g(y) \leq q \Rightarrow f(y) \leq f(x).$$

The *limit function* associated with the problem is

$$F(z) = \text{Sup}\ \ [f(x) : g(x) \leq z],$$

and a *support solution* u is defined by the condition

$$D(u) =:\ \ F(z) - F(q) \leq u(z - q)\ \ \text{for all}\ \ z,$$

equivalent to u being a support gradient of F at the point $z = q$.

Support solutions correspond to Lagrange multipliers in that they are variables associated with the constraints that give a means for characterizing optimal solutions. Thus, for a pair (x, u), *complementary slackness* is defined by

$$C(x,u) =: g(x) \leq q,\ \ u \geq 0,\ \ ug(x) = uq,$$

and a *shadow solution* by

$$\begin{aligned} S(x,u) &=: f(x) - ug(x) \\ &\geq f(y) - ug(y) \qquad \text{for all}\, y. \end{aligned}$$

An important proposition, not requiring any assumptions whatsoever about the set A or the functions f and g defined in it, is that any given pair (x, u) is a shadow solution with complementary slackness if and only if x is an optimal solution and u a support solution, that is,

$$M(x)\ \&\ D(u) \Leftrightarrow C(x,u)\ \&\ S(x,u).$$

For characterizing optimal solutions by means of the condition on the right, the outstanding issue therefore is the existence of a support solution. We

will find this guaranteed under conditions natural for economics at least.

A *convex problem* is one where f is a concave function and the elements of g are convex. The only importance is to make the limit function F concave. Then it has a linear support, and so a support gradient providing a support solution, at any interior point of the region where it is finite. Now with $F(q)$ finite, Slater's condition which requires $g(x) < q$ for some x assures that q is exactly such a point. Thus for a convex problem with Slater's condition, and with $F(q)$ finite, as it must be if an optimal solution exists, we do have the existence of a support solution, and so a characterization of all optimal solutions by means of shadow solutions with complementary slackness.

It is a short step from here to the characterization by means of Kuhn–Tucker conditions. These apply to a problem where the activity set A is a space of non-negative vectors, and the functions are differentiable. All that has to be known further is that for a differentiable concave function $\varphi(x)$ subject to $x \geq 0$ to be a maximum it is necessary and sufficient that

$$\varphi_x \leq 0, x \geq 0, \varphi_x x = 0.$$

Applied to the Lagrangian $f(x) - u[g(x) - q]$, with u fixed and non-negative, and x restricted non-negative, the conditions $S(x, u)$ for a shadow solution become

$$f_x - ug_x \leq 0, x \geq 0, \quad (f_x - \quad ug_x)x = 0,$$

and so now, with complementary slackness $C(x, u)$, we have the Kuhn–Tucker conditions. In case $x > 0$ the conditions just obtained reduce to $f_x = ug_x$, in other words, ordinary Lagrange conditions, the support solution u providing the multipliers.

With F concave, it is differentiable at a point q if and only if it has a unique support gradient u there, and then the support gradient coincides with the differential gradient, $u = F_q$. Thus uniqueness of support solutions is associated with differentiability of the limit function F at the point q. The identification $u = F_q$ that can be made in this case is comparable with the identity of classical Lagrange multipliers with derivatives of the stationary value function. But this new multiplier theory, even for the Kuhn–Tucker case, in no way depends on differentiability of the limit function. Also, for the linear approximation near q that is available in the differentiable case, we know more about it in that the error is always positive, or that it overestimates the limit function, not just locally but everywhere. Consider now a standard linear programming problem.

$$(M) \quad \text{Max} \quad px : ax \leq q, x \geq 0.$$

Another characterization for support solutions of LP problems can be noted, coming from the homogeneity. Thus, with F as the limit function of (M), the condition for u to be a support solution becomes

$$F(q) = uq, F(z) \leq uz \quad \text{for all} \quad z.$$

Since (M) is a convex problem the foregoing will apply to it. Also it has the required form for application of the Kuhn–Tucker conditions, which, following the way we put them before with some rearrangement in the second line, become

$$\begin{array}{llll} ax \leq q, & u \geq 0, & uax = uq, \\ ua \leq p, & x \geq 0, & uax = px. \end{array}$$

We know from the foregoing that (x, u) is a solution of these conditions if and only if x is an optimal solution and u a support solution of the problem (M).

There is a symmetry in the situation that enables these conditions to be read differently. With an exchange of role between x and u they become Kuhn–Tucker conditions for the problem

$$(W) \ \text{Min} \, uq : ua \geq p, u \geq 0,$$

and so they hold if and only if u is an optimal solution and x a support solution of (W). It follows that support solutions of either problem are identical with optimal solutions of the other.

Of course, (*M*) and (*W*) are a standard dual pair of LP problems, and so by the LP duality theorem one has an optimal solution if and only if the other does. Hence an LP problem has a support solution if and only if it has an optimal solution. Most remarkable is the way for finding support solutions for an LP problem, as it were differentiating the limit function or finding the 'Lagrange multipliers', by finding optimal solutions for another LP problem – and we know how to do that.

See Also

- ▶ Convex Programming
- ▶ Hamiltonians
- ▶ Non-linear Programming

Bibliography

Afriat, S.N. 1969. *The output limit function in general and convex programming and the theory of production*. In 36th National Meeting of the Operations Research Society of America, Miami Beach, Florida, November 1969. Reprinted, *Econometrica* 39(1971): 309–339.

Afriat, S.N. 1970. The progressive support method for convex programming. 7th Mathematical Programming Symposium, The Hague, 1970. *Journal of Numerical Analysis* 7 (3): 44–57.

Afriat, S.N. 1971. Theory of maxima and the method of Lagrange. *SIAM Journal of Applied Mathematics* 20: 343–357.

Afriat, S.N. 1986. *Logic of choice and economic theory*, Part V: Optimal programming. Oxford: Clarendon Press.

Dantzig, G. 1963. *Linear programming and extensions*. Princeton: Princeton University Press.

Kuhn, H.W., and A.W. Tucker. 1950. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Lagrange, J.L. 1762. Essai sur une nouvelle méthode pour determiner les maxima et minima des formules intégrales indéfinies. *Miscellanea Taurinensia* 2: 173–95. (Also *Théorie des fonctions analytiques,* 1797.)

Slater, M. 1950. *Lagrange multipliers revisited: a contribution to non-linear programming*. Cowles Commission Discussion Paper, Math. 403, November. University of Chicago.

Laissez-faire, Economists and

Roger E. Backhouse and Steven G. Medema

Abstract

This article traces economists's attitudes towards government intervention since the term 'laissez-faire' was first used in late 17th- or early 18th-century France. Understanding of the term has changed significantly since then. Adam Smith, popularly associated with laissez-faire, had a much more nuanced and pragmatic view of the role of the state, as did many of the classical economists and their neoclassical successors. Dissatisfaction with certain aspects of industrial capitalism led to a more interventionist stance during the 20th century, though the second half of the century saw something of a reversion towards the classical approach.

Keywords

Banking School; Bastiat, C. F.; Behavioural economics; Boisguilbert, P.; Bright, J.; Cairnes, J. E.; Cambridge School; Capitalism; Chalmers, T.; Chicago School; Cobden, R.; Consumer surplus; Corporatism; Cowles Commission; Currency competition; Currency school; Division of labour; Economic freedom; Efficient allocation; Free banking; Gold standard; Government failure; Great depression; Hayek, F. A. von; Hobson, J. A.; Information, economic of; Institutionalism; Keynes, J. M.; Keynesianism; Laissez-faire; Lange, O. R.; List, F.; Manchester school; Marginal revolution; Market failure; Market socialism; Markets; Marshall, A.; McCulloch, J. R.; Mercantilism; Mill, J. S.; Minimum wages; Mises, L. E. von; Mont Pèlerin Society; Natural liberty; New classical macroeconomics; New Deal; North, D.; Optimal resource allocation; Planning; Public choice; Rational behaviour; Rational choice; Ricardo, D.; Role of government; Samuelson, P. A.; Sidgwick, H.;

Smith, A.; Socialist calculation debate; Underconsumptionism; Utilitarianism; Viner, J.; Walras, L.

JEL Classifications
B1

Origins

The maxim 'laissez-faire' is commonly attributed to Vincent de Gournay (1712–1759), on the basis of a claim made the Physiocrat Du Pont de Nemours. However, his likely source, Turgot's 'Eloge de Gournay', did not attribute the phrase to Gournay, but implied that Gournay agreed with a well-known remark, made to Louis XlV's minister, Colbert, 'laissez-nous faire'. This remark was apparently made around 1680 by François Legendre, a merchant and author of a text on commercial mathematics, and may well have been an unpremeditated answer to a question (Castelot 1987). However, Legendre's contemporary, Pierre de Boisguilbert (1646–1714), repeatedly used the phrase 'laisse faire la nature' ('leave nature alone'), arguing that interference in business spoiled everything, even if it was well-intentioned (Faccarello 2000). The same ideas appear in English writings of the same period, though the phrase itself does not occur in the writings of commentators such as Nicholas Barbon and Dudley North, writing in the early 1690s, and Henry Martyn and Bernard Mandeville in the early 1700s. North (1691, p. 37), for example, wrote, 'no people ever yet grew rich by policies; but it is peace, industry, and freedom that brings trade and wealth, and nothing else.' By the mid-18th century, the idea of laissez-faire was well known, perhaps most clearly stated by the Marquis d'Argenson, in 1858: 'Laissez faire ought to be the motto of every public authority' (Castelot 1987; see also Oncken 1886).

It was Adam Smith who became associated, more than any other economist, with laissez-faire during the 19th and 20th centuries, even though he neither invented the idea, nor used the phrase. The elements from which his *Wealth of Nations* was constructed may not have been original, but the vision of society that he presented, with its emphasis on natural liberty, resonated widely. Smith used the idea of liberty as a radical idea that, though cautiously expressed, placed him alongside radicals such as Tom Paine and Condorcet. Liberty had a political as well as an economic dimension, involving freedom from being oppressed by guilds and monopolies as much as freedom from government interference in one's affairs. However, to those for whom such talk of liberty smacked of Jacobinism and the threat to property posed by the French Revolution, Smith could be reinterpreted as advocating a narrower economic freedom, more conservative in its political implications. Such a reinterpretation happened within a decade of his death (Rothschild 2001).

The Case for Laissez-faire

Smith's case for the market did not rest on any claim that it would produce an optimal allocation of resources. Instead, he argued that the system of natural liberty would produce a better outcome than would intervention by the state. There were hints concerning efficient allocation of resources, as on the only occasion when he used the phrase 'invisible hand' in the *Wealth of Nations*: in seeking his own advantage, 'every individual necessarily labours to render the annual revenue of the society as great as he can' (Smith 1776, p. 456). Smith opposed mercantilist policies so strongly, not because they prevented an efficient or optimal allocation of resources, or because state action was inherently less efficient than private, but because mercantilist policies were typically the result of using state power to serve the interests of a privileged minority. Merchants conspired to restrain trade, using the state where they could. Smith supported laissez-faire because removal of mercantilist restrictions on trade would help to undermine monopoly, enabling individuals to bring their capital into competition with those

who were earning high profits and allowing labour to flow freely between industries and regions. But Smith's support for laissez-faire was not for laissez-faire *in vacuo*: his system presumed a framework of justice and morality, the basis for which he had analysed in his *Theory of Moral Sentiments* (1759), a book to which he continued to attach great importance, revising it long after the *Wealth of* Nations was published, and in his lectures on jurisprudence delivered in the 1760s (1978).

The classical economists' case for laissez-faire was substantially Smithian, though more narrowly focused on economic freedom. Their consumption-oriented view led them to the belief that freedom of choice was desirable for consumers, and that freedom for producers was the most effective means of satisfying these consumer desires. It was thought that the impersonal forces of the market, working through the system of natural liberty, would then serve to harmonize these interests – or at least would do so to a greater and more beneficial extent than would other systems. The case was comprised of some arguments, such as David Ricardo's theory of international trade, that could be interpreted in terms of optimal resource allocation, but it centred on raising the growth rate. However, some economists saw the case for laissez-faire as primarily a moral one, linked to arguments from evangelical Christian theology. Where Ricardo and many other economists focused on the link between laissez-faire and economic growth, economists such as Thomas Chalmers endorsed laissez-faire because it allowed individualistic capitalism to have its full educative, retributive and purgative effects. There has even been debate over whether this moral case for laissez-faire was in practice more influential than the economic one (cf. Hilton 1988; Gash 1989). Certainly in America during this period, the belief in laissez-faire could not be separated from the Protestant spirit of the times, and a belief in its virtues was considered a necessary identifying mark of an economist. When it came to free trade, there was the additional dimension, emphasized by John Bright and Richard Cobden, arguably the most influential advocates of laissez-faire in Victorian Britain, that free international commerce held out the prospect of harmony between nations.

The most outspoken supporter of laissez-faire, however, was probably the French writer Frederic Bastiat, a brilliant economic journalist whose vivid examples (for example, candle-makers petitioning for protection against unfair competition from the sun) were influential in making the case for free trade. Standing in a French laissez-faire tradition going back to the 18th century, he linked laissez-faire with harmony between classes, in contrast with the class conflict seen by many English economists. In the United States, laissez-faire was also more than simply an economic doctrine, as is shown by the implications of the slogan of 'free labour' in a society divided over slavery. Along with the sanctity of private property it was part of a moral order that was believed to produce a harmonious society: free enterprise was strongly associated with the virtues of hard work and republican democracy (see ▶ United States, Economics in (1776–1885) and ▶ United States, Economics in (1885–1945)).

It was only towards the end of the century, with the developments commonly known as the marginal revolution, that economists began to argue that free competition might produce an optimal allocation of resources, thereby opening up a new defence of laissez-faire. Léon Walras (1954, p. 255) showed that if two conditions – that each product had only one price in the market and that prices equal corresponding costs of production – were satisfied, free competition would produce 'the greatest possible satisfaction of wants'. Marshall offered a doctrine of 'maximum satisfaction' that wedded his demand–supply apparatus with the concept of consumer surplus. However, they did not use these arguments to make a case for laissez faire, for their arguments showed much more clearly than did those of their predecessors why laissez-faire might in practice fail to produce such an optimal allocation. For example, immediately after stating his theorem, Walras pointed out that economists typically exaggerated the implications of the principle of laissez-faire: the conditions of a uniform price and equality of price and cost of production would often not be

satisfied and, in any case, it the theorem did not apply to the question of property.

The Limits to Laissez-faire

Viner (1960, p. 45), in one of the classic studies of the history of laissez-faire, started by saying that he understood laissez-faire to mean:

> the limitation of government activity to the enforcement of peace and of 'justice' in the restricted sense of 'commutative justice,' [justice in exchange] to defense against foreign enemies, and to public works regarded as essential and as impossible or highly improbably of establishment by private enterprise or, for special reasons, unsuitable to be left to private operation.

However, whilst Viner is correct to argue that laissez-faire did not mean anarchy and a complete absence of government intervention, his definition begs the question of how much intervention should be allowed: of what are the limits to laissez-faire.

Smith's view of the role of government is close to Viner's view of laissez-faire. The duties of the sovereign included maintaining justice, police, defence and such beneficial public works as would not otherwise be provided. This included support for transport and education – both of which Smith thought essential contributors to the wealth of a nation. It is important to note, though, that Smith's conception often went beyond modern views. For example, his support for education and for the arts was grounded in part in his concerns about the stultifying effects of the division of labour. His analysis of national defence led him to advocate a standing army rather than a militia because of a concern about the problems of attracting the right sort of people to military service in an increasingly wealthy commercial society. Smith's view of the appropriate sphere for state action also went significantly beyond the traditional public goods categories. He supported regulations dealing with public hygiene, legal ceilings on interest rates (to prevent excessive flows of financial capital into high-risk ventures), light duties on imports of manufactured goods, the mandating of quality certifications on linen and plate, certain banking and currency regulations to promote a stable monetary system, and the discouragement of the spread of drinking establishments through taxes on liquor (this being one of various regulations Smith advocated to compensate for the imperfect knowledge – or diminished telescopic faculty – of individuals). He also argued for measures that came within what Viner described as commutative justice. For example, he supported regulations that restricted wages in the interests of the labourer (that is, minimum wages) on the grounds that these redressed the imbalance between worker and employer.

The 19th-century classical economists, while holding to a strong belief in the market as an allocation mechanism, also believed that the market could only operate satisfactorily – harmonizing actions of self-interested agents with the interests of society as a whole – within a framework of legal, political, and moral measures that facilitated certain forms of action while restricting others. They were, in essence, pragmatic reformers, inclined towards laissez-faire but in practice willing to consider each case on its merits. We see this reflected in John Ramsay McCulloch's assertion in 1848 that 'The principle of *laisser-faire* may be safely trusted to do in some things but in many more it is wholly inapplicable; and to appeal to it on all occasions savours more of the policy of a parrot than of a statesman or a philosopher' (McCulloch, quoted in Robbins 1952, p. 43). Over two decades later, John Elliot Cairnes (1870, p. 244) was even more forthright, asserting that the maxim of laissez-faire had 'no scientific basis whatever' but was a 'mere handy rule of practice'. In terms of specific policies, they were willing to support an increasing range of interventions from factory legislation to the state provision of education, the poor laws and measures to promote public health (Robbins 1952; O'Brien 2004).

However, whilst the classical economists, like Smith, saw many cases where government action could improve on what would result from laissez-faire, they remained suspicious of government and were vociferously opposed to policies – like those of mercantilism, but also many others – that they believed served the interests of particular

groups at the expense of the larger population. They were optimistic that the insights of political economy could be used to point the discipline in a direction that would be beneficial to society and help mitigate the negative effects of partisan advocacy within that process (for example, Mill 1859, 1861, 1862).

More radical objections to laissez-faire were found outside Britain. The name 'Manchesterism' was widely used, particularly in Germany, to denote British laissez-faire doctrines, and was allegedly the ideology of Manchester's manufacturing classes. The most penetrating critique came from Friedrich List in *The National System of Political Economy* (1856). As Britain had industrialized first, free trade was in her interests, because other countries could not compete; until they were in a position to do so, tariffs were needed. List's ideas were particularly influential in the United States, where economists such as Henry Carey were able to combine commitment to individualism and free enterprise with support for protective tariffs.

One of the additional elements introduced after Smith was the utilitarianism of Bentham and his followers. Though utilitarianism has, on account of the prominence of Philosophic Radicals within political economy, been equated with laissez-faire individualism, this is not correct. On the one hand, there was an authoritarian streak in utilitarianism, from Bentham to reformers such as Edwin Chadwick. On the other hand, there were many supporters of laissez-faire, of whom Gladstone is perhaps the outstanding example, alongside many evangelical political economists, who would have no truck with Benthamite anti-religious rationalism.

The trend away from laissez-faire has its roots in the utilitarian tradition, for utilitarianism provided a basis on which exceptions could be justified. John Stuart Mill (1848, Book V, ch. XI), in what became the dominant textbook on political economy, laid out an extensive list of cases where the system of natural liberty failed to generate outcomes in the best interests of society. He argued that government interference was justified when individuals' actions had spillover effects on others, when individuals did not have the capacity properly to judge the consequences of their own actions or when what would now be called principal-agent problems were present. Prominent here, too, was the distribution question: the classical period witnessed increasing concern about poverty but saw attempts at reducing it as at best futile (owing to natural laws governing distribution) and possibly even counterproductive (because redistributive measures could exacerbate the population problem). Mill challenged the received view here by positing that the laws of distribution were, in fact, mutable, and that state action had the potential to significantly improve the lot of the poor. However, his starting point remained the maxim that '*Laisser-faire*, in short, should be the general practice: every departure from it, unless required by some great good, is a certain evil' (Mill 1965, p. 945). It was not just Mill who used utilitarianism as a means of justifying departures from laissez-faire. Robert Lowe, a controversial Liberal politician, at one time Chancellor of the Exchequer, had very much a Smithian view of the merits of laissez-faire but used utilitarian arguments to justify an increasing number of exceptions to this rule (Maloney 2005). William Stanley Jevons sought to move economic theory sharply away from the framework laid down by Ricardo and Mill, but used utilitarian arguments to justify extensive state intervention.

These utilitarian defences of state intervention were part of a much broader move away from laissez-faire from around the 1870s and 1880s when there developed widespread consciousness of what was, in Britain, called 'the social problem' at a time when the electorate in many European countries was widening to include the members of the working class (see Hutchison 1953, 1978). One reason for the timing was the long recession that followed the collapse of the worldwide boom of 1873 and the severe, and in some countries prolonged, unemployment that resulted. Questioning of laissez-faire was particularly strong in Germany, where the Verein für Sozialpolitik was founded, essentially as an interventionist think tank (see ▶ Historical School, German). Its members, of whom Gustav Schmoller was preeminent, were known as the 'Socialists of the lectern'. These attitudes carried

over to the United States: many of the founders of the American Economic Association were exposed to them whilst taking their doctorates in Germany, only to find, on their return, a conflict with traditional laissez-faire attitudes. Their challenge to laissez-faire affected not just economic analysis, but economic policy: in the United States, the rise of big business was associated with the development of numerous and very obvious anti-competitive practices, which resulted in the government developing policies of industrial regulation not found in other countries, at least in relation to inter-state trade, culminating in the anti-trust acts of 1890 and 1913. Economists supported such measures with analysis of phenomena such as 'cut-throat' competition that went beyond anything found in, say, Jevons, Walras or Marshall (see ▶ United States, Economics in (1885–1945)).

The British approach was dominated by the Cambridge School, at the headwaters of which was Henry Sidgwick, the author of one of the classic defences of utilitarianism ethics (1907). Sidgwick (1904) took Mill's analysis further: *all* outcomes that constituted departures from social utility maxima were potential candidates for government intervention. Sidgwick's optimism about the prospects for state action marked a significant turn. He was convinced that recent reforms in governance structures – such as the establishment of boards and commissions staffed by experts – portended great things for the ability of state action to improve on market performance.

Sidgwick's perspective signalled what was to become a distinctive Cambridge approach to issues of laissez-faire, continued by Marshall (1890) and A.C. Pigou (1912, 1920). Marshall wedded his demand–supply analysis with the concept of consumer's surplus to provide a tool with which the welfare implications of laissez-faire and government intervention could be analysed. In analysis since seen as flawed through its neglect of producer's surplus, Marshall argued that subsidies to industries characterized by increasing returns and taxes on industries operating under decreasing returns could enhance efficiency. Pigou took all this a step further with his analysis of private and social net products, which proved to be a very effective tool for illustrating both the nature of market failures and the means by which government corrective actions could prod markets toward efficiency. He argued that divergences between private and social net products constituted a '*prima facie* case' for government intervention, but he also allowed that the state will not necessarily be capable of improving on market performance. Like his predecessors, Pigou was optimistic that governmental reforms held great promise, but he was also concerned about many of the governance problems that we now associate with public choice analysis. The policy conclusions of the Cambridge economists, including the case for free trade, rested as much on beliefs about the competence of government to implement beneficial policies as on the results of formal economic theory.

The First World War and Its Aftermath

Laissez-faire was far from universally accepted before the First World War, but the move towards the welfare state and towards regulation of industry was generally gradual (free trade had never become universal, some countries never having abandoned protective tariffs). Economists made frequent concessions to socialism (this was easy because the term had such an elastic meaning) but could maintain the idea that laissez-faire should remain the general rule. After 1918, that confidence was harder to maintain. The Bolshevik revolution and the establishment of the Union of Soviet Socialist Republics (USSR) presented the challenge of an alternative economic system. Economic dislocation was widespread in Europe in the 1920s and worldwide after the onset of the Great Depression. To an extent unparalleled before 1914, laissez-faire and even capitalism were called into question, in the writings of economists as much as among politicians and policymakers.

Of particular significance was the extension of discussions of laissez-faire to what would now be considered macroeconomic issues – money and the business cycle. 'Free banking' might exist in some American states, but the need for some sort

of monetary policy had been generally accepted since the bullion debates during the Napoleonic Wars. Though there were exceptions, it became accepted that paper money should have a fixed value in terms of precious metal. There were several reasons why this was seen as consistent with laissez-faire. To allow the value of paper to fall below par was to defraud those who had entered into contracts denominated in terms of money. A metallic standard, which increasingly meant the gold standard, facilitated trade. Most important, though there were underconsumptionists (more than are often recognized), they were in a clear minority among economists.

The parallel with 20th-century debates over laissez-faire in macroeconomics is found in the debate between the Currency and Banking Schools in the 1840s (see ▶ Banking School, Currency School, Free Banking School). The Currency School sought to prevent the emergence of financial crises by making paper currency behave like a metallic one, removing discretion from central bankers. In contrast, the Banking School argued that, in times of depression, a central bank should pursue an accommodating monetary policy, lending according to sound banking principles. Strictly speaking, this was a debate about the type of policy to be pursued, not whether or not to intervene, but it posed the issue of discretion in monetary policy that came to be associated, in the 20th century, with debates over laissez-faire. Such ideas framed much of the discussions of central bank policy as late as the inter-war period, when the appropriate policy for the US Federal Reserve system was being debated (Laidler 1999, chs 8–9).

The extent to which such a way of thinking carried over into the 20th century is illustrated by the 'Austrian' theories of money. Though in general ardent supporters of laissez-faire, Ludwig von Mises and Friedrich Hayek argued for the implementation of what they considered appropriate monetary policy. Mises (1912, pp. 456–63) supported the gold standard on the grounds that it rendered the value of money independent of political influence. Management of the currency meant inflation, a policy inevitably doomed to eventual failure. Hayek, though theoretically innovative, maintained this emphasis on sound, or 'neutral' money; the problem of the business cycle was caused by the supply of money being too elastic. Despite his otherwise impeccable credentials as a supporter of laissez-faire, it was only in the 1970s that he turned to completely free banking and competition in the supply of currency (Hayek 1999).

Though his target was the British authorities, this was the mindset that John Maynard Keynes attacked in his *Tract on Monetary Reform* (1923). He argued that to regard the gold standard as a fact of nature was to perpetuate an illusion. 'There is,' he wrote, 'no escape from a "managed" currency, whether we wish it or not' (Keynes 1971, p. 136). He continued (1971, p. 138):

> A regulated non-metallic standard has slipped in unnoticed. *It exists*. Whilst the economists dozed, the academic dream of a hundred years, doffing its cap and gown, clad in paper rags, has crept into the real world by means of the bad fairies – always so much more potent than the good – the wicked ministers of finance.

It was but a short step from this to announcing '[t]he end of laissez-faire' (Keynes 1972, pp. 272–94). His account of laissez-faire focuses on the philosophical and political rather than the economic, his point being that it cannot rest on ideas of natural liberty, for there is no such thing. It was necessary, he argued, to work out the agenda and non-agenda of the state without the Benthamite prior assumption that interference was likely to be 'generally pernicious' (1972, p. 288). The agenda for the state should comprise those things that are otherwise not done, which he identified as regulation of currency and credit, management of investment, and policy in relation to population size (1926, p. 292). His *General Theory of Employment, Interest and Money* (1936) provided a new theoretical justification for such ideas, but the idea that the state's main agenda item was the maintenance of the level of investment, remained.

Most of Keynes's arguments were far from novel. J.A. Hobson and other underconsumptionists had long questioned the ability of unregulated capitalism to produce the appropriate level of saving. Not only had it been argued, even

before 1914, that government spending could raise the level of employment, but schemes for doing so had been worked out. The significance of his arguments, which became clear only from the 1940s, lay in the fact that an economist at the heart of the establishment was arguing against laissez-faire from a macroeconomic point of view. Furthermore, his attack on the philosophical foundations of laissez-faire, by someone who was far from being a socialist, indicated a changing climate of opinion towards one in which management of the economy came to be seen as a central role of government.

Planning was also becoming more acceptable in the United States throughout the inter-war period. It has been argued that this marked a radical departure from previous attempts to reform the economy because it was 'predicated on the assumption that intervention . . . was necessary for a well-functioning, dynamic economy' (Balisciano 1998, p. 154; see also Barber 1985, 1996). The First World War had shown that planning could raise output above what had been thought possible, and economic fluctuations in the immediate post-war period suggested that government intervention might be desirable. There was a move to create a new economics, appropriate to a new age, exemplified in the White House by Herbert Hoover, an engineer who turned readily to experts. The move towards a scientific economics that could perform this task was represented by institutionalism (see ▶ Institutionalism, Old) but planning took many forms, from the social planning of Rexford Tugwell and John Maurice Clark to the macroeconomic planning of Laughlin Currie (see Balisciano 1998). The move towards providing a scientific foundation for policy extended beyond institutionalism: for example, both Wesley Mitchell and Irving Fisher called for quantitative research. These various strands of thought came together in the New Deal, with its mixture of microeconomic planning, macroeconomic management and extensive quantitative research.

In continental Europe, planning was observed in the Soviet Union and in Germany under National Socialism. In other countries, corporatist ideas were highly influential. When placed alongside experience of the Great Depression, this raised the question of whether capitalism itself, let alone laissez-faire, was a viable alternative to planning. The socialist calculation debate, initiated by Otto Neurath and von Mises immediately after the First World War, tackled the question of whether a planned economy could be as efficient as a capitalist one. The significance of this controversy is twofold. In making the case that it was theoretically possible to design a socialist economy that was as efficient as a capitalist one, Oskar Lange and the so-called market socialists were shifting the climate of opinion in favour of planning. However, perhaps more significant in the longer term is the fact that planning was defended using arguments about the optimality of a perfectly competitive equilibrium. This took the arguments of Walras and Marshall a stage further, towards the post-war welfare theorems of Kenneth Arrow and Gerard Debreu. A defence of socialism could, with a small twist, be turned into an argument for laissez-faire.

The most theoretically innovative critic of the central planners was Hayek, who developed the idea that the market could be seen as an information-processing mechanism (Hayek 1937; see also Gamble 2006). The information possessed by modern societies was necessarily limited, imperfect and dispersed among many individuals, so to assume, as did the market socialists, that this knowledge could be available to central planners was a mistake. Markets enabled prices and economic activities to reflect the knowledge held by millions of distinct individuals and organizations. The significance of this theory is that it reinforces the point that arguments for laissez-faire do not need to rest on any claim that it produces an optimal outcome. If knowledge is imperfect, as Hayek claimed, it is not meaningful to argue in terms of optimality.

The Second World War and After

During the Second World War, planning was widely practised, not just in Germany and the Soviet Union but also in Britain and the United States, perhaps inevitably when military uses

accounted for around 40 per cent of national production. Unlike in the First World War, careful attempts were made to plan for the post-war order and although this was to be a liberal world order, based on free trade and free movement of capital, it was to be a planned order, with appropriate national and international institutions to support it. Experience of the First World War was taken as demonstrating that a well-functioning free market economy would not occur spontaneously. The degree and nature of planning and commitment to laissez-faire varied from country to country: the United States may have been at one end of the spectrum, with suspicion that planning might be tainted by Communism, and with government accounting for a lower share of national output than in Europe, but the importance of the defence sector during the Cold War meant that the role of government was far-reaching. Though there was a retreat from the level of planning achieved during the war, and even compared with the New Deal, in favour of a free market economy, government remained very significant.

Economics had also changed, becoming more technical, more mathematical (see, for example, ► United States, Economics in (1945 to Present)). However, the relationship of this change to thinking about laissez-faire was complex. Many of the techniques used in this more technical economics had roots in economists' wartime activities, and were linked with planning. The Cowles Commission, the main centre of mathematical economics in the 1940s, was closely associated with these developments and was also linked, through Oskar Lange, with socialism. A case can be made that microeconomic theory in this period strengthened the case against laissez-faire by developing theories of market failure. General equilibrium theory may have been seen by outsiders as demonstrating rigorously the efficiency of competitive markets, but the restrictive assumptions needed could equally be taken as demonstrating that an efficient allocation of resources required conditions that could never be satisfied in the real world.

It was in macroeconomics that the challenge to laissez-faire was strongest, the nearest to a consensus view being the neoclassical synthesis articulated in the third edition of Paul Samuelson's *Economics* (1955). This proposed that if demand management could maintain full employment, the allocation of resources between economic activities could be undertaken by the market. Laissez-faire was rejected at the macroeconomic level in favour of a 'Keynesian' policy of demand management (see ► Keynesianism). At a microeconomic level, laissez-faire was limited by the need to provide public goods, deal with externalities and control monopoly. This left much scope for debate over precisely where the limits to laissez-faire lay, from those who favoured extensive intervention to the Chicago School, which challenged the need for active competition policies and, increasingly, the Keynesian consensus.

The pervasiveness of planning in the late 1930s and early 1940s provoked a response from some scholars who believed that classical liberal values were threatened. The most prominent such response was by Hayek, whose *The Road to Serfdom* (1944) became a best seller. In 1947 he helped establish the Mont Pèlerin Society, which became the centre of a network of economists committed to free-market ideas. This network encompassed research institutes aimed at influencing policy and academic economists, of which the most significant was a group centred on Chicago. This offered a much more optimistic, and even radical, view of what could be achieved under laissez-faire than was generally accepted by economists in the 1950s and 1960s. Laissez-faire was as much an end as a means, exemplified in Milton and Rose Friedman's *Free to Choose* (1979).

The 1960s and 1970s saw the beginnings of a major shift in the way that economists approached issues related to laissez-faire. At the heart of this shift was an extension in the scope of the theory of rational choice to the point where it could encompass all aspects of behaviour (see ► Rationality, History of the Concept). Two developments were particularly important in moving economists towards laissez-faire. The first was the application of rational choice theory to government and bureaucracies, resulting in the development of a theory of government failure to parallel the earlier theory of market failure. Rent-seeking, legislative vote trading and bureaucratic waste took their

places alongside externalities and public goods as phenomena to be taken into account. This was most visible in public choice theory, but spread much more widely. The second was the transformation of macroeconomics associated with the new classical macroeconomics. Rational behaviour was taken to imply that markets would clear and that agents would form expectations rationally, which led to a presumption that attempts to stabilize economic activity would be counterproductive; that laissez-faire applied at the macroeconomic level. This was believed to explain the apparent breakdown of Keynesian policies in the 1970s. This did not go unchallenged, but there was a clear shift in the weight of economists' opinions on laissez-faire at both microeconomic and macroeconomic levels.

However, other developments worked in the opposite direction. There was much work on the economics of information, which added to the weight of the evidence for why free markets might not be efficient. These involved questioning some of the most basic ideas of supply and demand on which much of the traditional case for laissez-faire rested, a point made most forcefully by Joseph Stiglitz. Market failures can occur in both the production and dissemination of information due to the informational asymmetries and uncertainty that result. A lack of effective futures markets causes intertemporal inefficiencies (for example, on the environmental front), moral hazard and adverse selection problems can cause insurance markets to fail, and the use of education as a signalling and screening device can lead to overinvestment in education. At the macroeconomic level, problems associated with risk and information can cause financial markets to react in ways that are destabilizing. Game theory, too, presented problems, showing how strategic behaviour had a propensity to generate market outcomes that departed – sometimes substantially – from the dictates of optimality.

By the new millennium, some of the assumptions underlying this resurgence of laissez-faire thinking were being challenged. A form of Keynesianism re-emerged in the form of inflation targeting through interest rates, a development that reflected both macroeconomic theory and lessons learned from experience. Behavioural economics raised questions about human motivation and opened up the possibility of new ways of analysing economic behaviour. It is, however, too soon to tell what the implications of this will be for attitudes towards laissez-faire.

However, despite the resurgence of laissez-faire thinking, the context is radically different from that prevailing at the beginning of the 20th, let alone the 19th, century. In macroeconomics, the case for central banks operating according to rules so as to stabilize economic activity is, in some sense, almost universally accepted. Debates centre on what those rules should be, not whether there should be rules. At the micro level, there has been a significant expansion in the sphere of market activity since the 1980s, as a result of the deliberate creation of new markets, from financial options to CO_2 emissions. These markets are not simply heavily regulated: many of them are designed by government, usually on the basis of economists' advice. Furthermore, the scale of government is now such that government contracts are an inherent part of the activities of many businesses. In such an environment, it can be questioned whether the traditional distinction between laissez-faire and government intervention has become out of date.

A further complication in discussions of laissez-faire results from the enormous expansion of international organizations, from the International Monetary Fund (IMF) and the World Bank to the World Trade Organisation (WTO) and the United Nations. These have made it meaningful to discuss alternatives to laissez-faire at an international level at the same time that the so-called globalization of economic activity has raised new questions about its benefits and costs to different groups. If trade is to take place within rules laid down by organizations such as the WTO and the IMF, should these rules allow governments to protect industries or workers from what they perceive to be unfair international competition? Does laissez-faire apply to national governments or simply to private organizations? This is a complicated question in a world where many private companies are substantially shaped by their relations with governments.

Conclusions

It has been argued that the developments of recent decades have taken us back to Adam Smith and a laissez-faire welfare economics asserting the efficacy of the market in channelling individual self-interest towards actions that benefit society; and to a pre-Keynesian era when the need for active macroeconomic management was not recognized. But this is not accurate. The 19th- and early 20th-century exponents of laissez-faire, from Mill to Pigou (and perhaps even to Lange or Samuelson), saw an ever-widening range of exceptions to the general rule. Their policy prescriptions reflected well-articulated ideas about market failure and much less completely theorized views about the capacity of government to remedy such problems. As a result of recent developments there is, in general, awareness that neither the market nor government is perfect – that the choice is between two highly imperfect alternatives. Theory cannot settle the matter unless reasons are adduced to play down the importance of market failure (the 'libertarian' response) or government failure (the 'socialist' response). Because of this, and because of the transformed role of government, there is a strong case for arguing that notions such as 'laissez-faire' and 'state action', especially if this is seen as an either/or choice, are not particularly helpful. However, there is a reversion to Smithian ideas in one sense: economists increasingly recognize, as did Smith, that markets do not exist apart from an institutional structure that includes the state and its legal system. Discussions of state action are not usually about *replacing* the market; rather, they are about nudging markets this way or that in order to obtain a more desirable outcome than would obtain otherwise.

See Also

- ▶ Banking School, Currency School, Free Banking School
- ▶ Historical School, German
- ▶ Institutionalism, Old
- ▶ Keynesianism
- ▶ Marginal Revolution
- ▶ Rationality, History of the Concept
- ▶ United States, Economics in (1776–1885)
- ▶ United States, Economics in (1885–1945)

Bibliography

Balisciano, M.L. 1998. Hope for America: American notions of economic planning between pluralism and neoclassicism. *History of Political Economy* 30(Suppl): 153–78.

Barber, W.J. 1985. *From new era to new deal: Herbert Hoover, the economists, and American economic policy, 1921–1933*. Cambridge: Cambridge University Press.

Barber, W.J. 1996. *Designs within disorder: Franklin D. Roosevelt, the economists, and the shaping of American economic policy, 1933–1945*. Cambridge: Cambridge University Press.

Cairnes, J.E. 1870. Political economy and laissez-faire. In *Essays on political economy*. London: Macmillan, 1873.

Castelot, E. 1987. Laissez-faire, laissez-passer: History of the maxim. In *The New Palgrave: A dictionary of economics*, vol. 3, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Faccarello, G. 2000. *Foundations of laissez-faire: The economics of Pierre de Biosguilbert*. London: Routledge.

Friedman, M., and R. Friedman. 1979. *Free to choose*. New York: Harcourt Brace.

Gamble, A. 2006. Hayek on knowledge, economics and society. In *The Cambridge companion to Hayek*, ed. E. Feser. Cambridge: Cambridge University Press.

Gash, N. 1989. Review of Hilton 1988. *English Historical Review* 104: 136–40.

Hayek, F.A. 1937. Economics and knowledge. *Economica N.S.* 4(13): 33–54.

Hayek, F.A. 1944. *The road to Serfdom*. London: Routledge.

Hayek, F.A. 1999. Good money. In *The collected works of F.A. Hayek*, ed. S. Kresge. London: Routledge.

Hilton, B. 1988. *The age of atonement: The influence of evangelicalism on social and economic thought, 1785–1865*. Oxford: Clarendon Press.

Hutchison, T.W. 1953. *A review of economic doctrines, 1870–1929*. Oxford: Oxford University Press.

Hutchison, T.W. 1978. *On revolutions and progress in economic knowledge*. Cambridge: Cambridge University Press.

Keynes, J.M. 1923. A tract on monetary reform. In *The collected works of John Maynard Keynes*, vol. 4. London: Macmillan, 1971.

Keynes, J.M. 1926. The end of laissez faire. In *Essays in persuasion. In the collected writings of John Maynard Keynes*, vol. 9. London: Macmillan, 1972.

Keynes, J.M. 1936. The general theory of employment, interest and money. In *The collected writings of John Maynard Keynes*, vol. 7. London: Macmillan, 1973.

Laidler, D.W. 1999. *Fabricating the Keynesian revolution: Studies of the inter-war literature on money, the cycle, and unemployment*. Cambridge: Cambridge University Press.
List, F. 1856. *The national system of political economy*. New York: Kelley, 1966.
Maloney, J. 2005. *The political economy of Robert Lowe*. London: Palgrave Macmillan.
Marshall, A. 1890. *The principles of economics*, 8th ed. London: Macmillan, 1920.
Mill, J.S. 1848. Principles of political economy, with some of their applications to social philosophy. In *The collected works of John Stuart Mill*, vol. 2–3, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.
Mill, J.S. 1859. *On liberty*. New York: Classics of Liberty Library, 1992.
Mill, J.S. 1861. Considerations on representative government. In *The collected works of John Stuart Mill: Essays on politics and society*, vol. 19, ed. J.M. Robson. Toronto: University of Toronto Press, 1977.
Mill, J.S. 1862. Centralisation. In *The collected works of John Stuart Mill: Essays on politics and society*, vol. 19, ed. J.M. Robson. Toronto: University of Toronto Press, 1977.
von Mises, L. 1912. *The theory of money and credit*. Trans. H.E. Bateson. Indianapolis: Liberty Classics, 1980.
North, D. 1691. *Discourses upon trade*. London: Thomas Basset.
O'Brien, D.P. 2004. *The classical economists revisited*. Princeton: Princeton University Press.
Oncken, A. 1886. *Die Maxime Laissez Faire et Laissez Passer*. Bern: Wyss.
Pigou, A.C. 1912. *Wealth and welfare*. London: Macmillan.
Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.
Robbins, L.C. 1952. *The theory of economic policy in English classical political economy*. London: Macmillan.
Rothschild, E. 2001. *Economic sentiments: Adam Smith, Condorcet and the enlightenment*. Cambridge, MA: Harvard University Press.
Samuels, W.J. 1966. *The classical theory of economic policy*. Cleveland: World.
Samuelson, P.A. 1955. *Economics*, 3rd ed. New York: McGraw Hill.
Sidgwick, H. 1904. *Principles of political economy*. London: Macmillan.
Sidgwick, H. 1907. *The methods of ethics*, 7th ed. London: Macmillan. Reprinted Indianapolis: Hackett, 1981.
Smith, A. 1759. The theory of moral sentiments. In *The works and correspondence of Adam Smith*, vol. 1, ed. D.D. Raphael and A.L. Macfie. Oxford: Oxford University Press, 1759.
Smith, A. 1776. An inquiry into the nature and causes of the wealth of nations. In *The works and correspondence of Adam Smith*, vol. 2, ed. R.H. Campbell, A. Skinner, and W.B. Todd. Oxford: Oxford University Press.
Smith, A. 1978. Lectures on jurisprudence. In *The works and correspondence of Adam Smith*, vol. 5, ed. R.L. Meek, P.G. Stein, and D.D. Raphael. Oxford: Oxford University Press, 1978.
Viner, J. 1960. An intellectual history of laissez faire. *Journal of Law and Economics* 3: 49–69.
Walras, L. 1954. *Elements of pure economics: Or the theory of social wealth*. Trans. W. Jaffe. Philadelphia: Orion.

Laissez-Faire, Laissez-Passer, History of the Maxim

E. Castelot

Gournay is still generally credited with being the inventor of this phrase, and this apparently on the authority of his friend Turgot, who, however, in his *Éloge* of Gournay, simply says:

> It ought to be added that the … system of M. de Gournay is remarkable in this respect, that … in all times and everywhere the desire of trade has been concentrated in these two words, *liberty* and *protection*, and most of all liberty. The remark of M. Legendre to M. Colbert is well known: '*laissez nous faire*' (Turgot, *Petite Bibliographie Economie*, p. 40, '*Il faut dire eoncre*', etc.).

This supposed agreement of the views of Gournay with the observation of Legendre has been translated by Dupont de Nemours into the positive statement: 'From his (Gournay's) profound observation of facts he had drawn the celebrated axiom, *laissez-faire*, *laissez-passer*' (*Oeuvres de Turgot*, ed. Daire, i. p. 258); and has been followed by most of the writers on economic literature down to M.G. Schelle, Dupont's last biographer (*Du Pont de Nemours et l'École Physiocratique*, Paris, 1888, p. 19).

In *Die Maxime Laissez Faire et Laissez Passer* (Bern, 1886), Professor August Oncken has thoroughly sifted and examined all available evidence on this subject, and comes to conclusions which may be definitively accepted, although he has not completely succeeded in identifying Legendre. The latter appears to have been François Legendre, the writer of an arithmetical treatise entitled

L'Arithmétique en sa Perfection selon l'usage des Financiers, *Banquiers et Marchands*, which went through nine editions between 1657 and 1687. Prof. Oncken has not been able to find out on what occasion the above reply was made to Colbert, but is inclined to believe that it must have been about 1680.

Still Legendre was a merchant and not a political writer; his answer was probably unpremeditated, and was wanting in the distinction of literary fame. In the writings of his contemporary, Boisguillebert, we meet, however, sentences which are closely allied to Legendre's utterance, such as: *Il n'y avait qu'à laisser faire la nature et la liberté* (*Factum de la France*, p. 286, ed. Daire), and, *Ainsi dans le Commerce de la Vie*, *elle* (nature) *a mis un tel ordre que pourvu qu'on laisser faire*, etc. (ibid., p. 280).

The worthy Norman magistrate, Boisguillebert, would thus have been the first to use with a scientific purpose, if not the actual first half of the maxim, at least language approaching to it. After him we must come down to the Marquis d'Argenson, in order to find a distinct and clear enunciation of the same principle conveyed still more pointedly in the essay to which he gave the title of *Pour gouverner mieux*, *il faudrait gouverner moins* (In order to govern better, we ought to govern less) (*Journal et Mémoires du Marquis d'Argenson*, 1858, vol. v). Here he emphatically declares that *Laissez faire*, *telle devrait être la devise de toute puissance publique* (*Laissez faire* ought to be the motto of every public authority), p. 364. The same line of reasoning is consistently followed, and similar expressions are used, in his *Pensées sur la Réformation de l'État* and in sundry contributions to the *Journal Économique*, the authorship of which has been brought home to D'Argenson (Oncken, *Die Maxime Laissez faire*, pp. 66–80). Neither Quesnay nor Adam Smith uses the expression, but it is printed several times in the *Ephémérides du Citoyen*, and now in its complete form (laissez-faire, laissez-passer), and constantly put into the mouth of Gournay (see quotations in Oncken, pp. 86–9). Mirabeau, Mercier de la Rivière, and Letrosne in their works give vent to the same theory, but under the parallel French or Italian form: *Le monde va de lui-même* or *Il mondo va da se* (The world goes by itself) (Oncken, pp. 84, 85). From what precedes, we may, it seems, safely conclude that if Gournay is not the actual inventor of the maxim, he put it into circulation through his conversations, after having contributed its second half.

Although the physiocrats had numerous contemporary adherents in Germany, the latter do not appear to have adopted the expression, unless the maxim of Iselin (born at Basle, 1728), *Lasset der Natur ihren Gang* (Let nature have her course), in his 'Ephemerids of the human kind' (*Ephemerids der menschheit*), be considered as an attempt towards a translation (Oncken, p. 127).

In England, J. Stuart Mill employed the actual French words *laisser faire* (but in the infinitive, not the imperative mood) in the table of contents of his *Principles of Political Economy*, as a heading to § 7 of ch. xi.

Bibliography

Boisguilbert, P.P.S. 1707. Factum de la France. In *Economistes et financiers du XVIII siècle*, ed. M.E. Daire. Paris, 1851.

Mill, J.S. 1848. *Principles of political economy*. London: J.W. Parker.

Oncken, A. 1886. *Die Maxime Laissez Faire et Laissez Passer*. Bern.

Schelle, M.G. 1888. *Du Pont de Nemours et l'école physiocratique*. Paris: Librairie Guillaumin et Cie.

Turgot, A.R.J. 1844. *Oeuvres de Turgot*. Ed. M.E. Daire, Paris. (Turgot's *Administration et oeuvres économiques*, ed. L. Robineau. Paris: Guillaumin, 1889, has the subtitle *Petit bibliotèque économique française et étrangère*.)

Lancaster, Kelvin John (1924–1999)

Ronald Findlay

Abstract

Kelvin Lancaster made at least three major original contributions to economic theory. The first, together with Richard G. Lipsey, is 'The General Theory of the Second Best' in the

area of welfare economics. The other was his 'characteristics' approach to the pure theory of consumer behaviour. The third, based on this new approach to consumer behavior was a solution to the problem of 'socially optimal product differentiation', which showed how to balance the consumer's desire for more variety in the choice of goods to consume against economies of scale in the production of each good.

Keywords

American Economic Association; Characteristics; Consumer choice; Econometric society; Johnson, H. G; Lancaster, K. J; Monopolistic competition; New trade theory; Pareto efficiency; Product differentiation; Scale economies

JEL Classification

B31

Kelvin Lancaster was born in Sydney, Australia, on 10 December 1924. He volunteered for the Royal Australian Air Force at the age of 18 and was trained as a bombardier. The war fortunately ended before this kindest and gentlest of men was required to release any bombs using the new Norden bombsight on which he had been trained. He graduated from the University of Sydney with a BSc in mathematics and geology (1948) and a BA (1949) an MA (1953), both in English literature. A growing interest in economics took him to the London School of Economics in 1953, where he obtained the BScEcon degree with First Class Honours as an external student without ever having taken a single course in economics, and his Ph.D. in 1958. He was on the faculty of the LSE from 1954 to1962, and immediately became one of the brightest stars of the famous seminar led since the early 1930s by Lionel Robbins, whose participants over the years included the likes of Hayek, Hicks, Kaldor, Lerner, Meade and many others of comparable stature.

Lancaster and Richard Lipsey, then also at the LSE, each submitted a paper to the *Review of Economic Studies*, edited by the indefatigable Harry Johnson, Lipsey's on tariffs and customs unions and Lancaster's on monopoly and nationalized industries. Johnson noted that they were both making the same general point, namely, that if one of the necessary conditions for a Pareto optimum failed to hold it was not in general desirable to make the remaining conditions hold. In other words, the Paretian conditions had to be fulfilled in their entirety for a 'first-best' optimum to be reached. If one condition failed to hold a 'second-best' optimum would in general involve departures from some or all of the others. Johnson suggested that the two papers be merged, making this fundamental general point and giving the customs union and nationalized industry problems as illustrative examples. The result was the celebrated paper on 'The General Theory of the Second Best' by Lipsey and Lancaster (1956) that has changed the way economists have since thought about economic policy in every field.

Lancaster moved to the United States in 1962, first to Johns Hopkins (1962–1966) and then to Columbia, following his wife Dvora, who had been admitted to Columbia Law School. He remained at Columbia for the rest of his career, becoming the John Bates Clark Professor of Economics in 1978. In 1966 he published 'A New Approach to Consumer Theory' in the *Journal of Political Economy*, following it in 1971 with a more detailed treatment in the book *Consumer Demand: A New Approach*. His attempt at a new approach to the classic problem of consumer choice was motivated by the desire to make this most parsimoniously elegant of all economic theories more operational and relevant to the modern industrial world of an almost infinite variety of products. The standard theory involved considering the consumer as maximizing a utility function $U(x)$ subject to a budget constraint $px = I$, where x is an n-dimensional vector of goods, p the corresponding vector of prices and I the income of the consumer. The basic idea of the alternative approach he proposed is to regard the arguments of the utility function not as goods but the *characteristics* or attributes of these goods that they provide to the consumer in varying amounts and proportions, the goods themselves being merely the means whereby the consumer satisfies his

essential wants. A simple version of the Lancaster approach therefore regards the consumer as maximizing $U(z)$, where z is an m-dimensional vector of characteristics, subject to $z = Bx$, where B is an $(m \times n)$ matrix representing the 'technology of consumption' or the amount of each characteristic embodied in each good, and the budget constraint $px = I$ as before. Lancaster regards the number of characteristics m as much smaller than the number of goods n in a modern economy.

Suppose, for purposes of illustration, that n is five and m is two. Given I and p we can find how much of each good can be obtained if all of the income is spent on that good alone. The amounts of each of these five goods obtained this way yield a pair of the amounts of the two characteristics that they embody. Number these goods from one to five in descending order of the ratio of the first characteristic to the second that they provide. Each of these five pairs can be plotted as a point in a diagram with the first characteristic on the vertical and the second on the horizontal axis. These points form the five vertices and the straight lines connecting them the four edges or flats of the 'characteristics-possibility frontier' or CPF available to the consumer, given his income and the prices of the goods that he is facing. Superimposing the map of convex indifference curves between the two characteristics specified by $U(z)$, we can find the optimal choice of the two characteristics for the consumer by the point at which the highest attainable indifference curve is tangent to the CPF. If the optimal point is on a flat the consumer will demand the convex combination of the two goods spanning the flat yielding that point; the only other possibility is for the optimal point to be at a vertex, in which case only the corresponding good is demanded. Each consumer will therefore demand at most two of the five goods available to him. Any other consumer will face the same price vector p for the goods and the same objective technology of consumption represented by the matrix B. Differences in income will result in radial expansions or contractions of the CPF, leaving its structure unchanged. The utility functions $U(z)$ of the consumers will all in general differ, leading to different choices of the at most two goods that each demands, so that each of the five goods will have a positive demand in the market as a whole if the tastes for characteristics are sufficiently diverse. Adding together the amounts of each good demanded by all the consumers, we obtain a point on the market demand function for that good at the given price vector p. Repeating the analysis described for all possible price vectors, we can generate the market demand functions for all five goods by the Lancaster method and then proceed as usual.

The power of this alternative approach is perhaps best revealed by the problem of new goods. In the standard theory we would have to recast the entire utility function U as a function of six instead of five arguments in our example, with almost no restrictions capable of being placed on the properties of the new function in comparison with the old. In the Lancaster model, however, the utility function in characteristics space $U(z)$ is entirely unaffected by the introduction of the new good. Given its price the new good will appear in the budget constraint with an additional sixth term and in the matrix B as an additional sixth column, leading to a sixth vertex and a fifth 'flat' for the new CPF. The new good thus leads only to a change in the CPF, which is common for all consumers, with all individual utility functions in the space of characteristics unchanged, instead of each having to be altered in its own particular way in the space of goods. By looking at the CPF we can see exactly which consumers will be affected and which not by the introduction of the new good. If we consider the cases of electric light and candles, automobiles and the horse and buggy, compact discs and vinyl LP records, it is clear that the new goods altered the technology of consumption for all consumers by providing a 'dominant' new good that drove out the competing old one on efficiency grounds, rather than leading to a simultaneous subjective shift in tastes by all consumers. Though developed for the analysis of consumer demand, the characteristics approach is also clearly applicable to portfolio selection between alternative financial assets, occupational choice problems in labour economics, provision of public goods and

services (see Lancaster 1991, Part 3) and many other areas.

The characteristics approach also led Lancaster naturally to the problem of 'socially optimal product differentiation' that he investigated initially in an article, Lancaster (1975), and with considerably more depth and detail in the 1979 major treatise entitled *Variety, Equity and Efficiency*. To explain the essentials of this problem, consider once again the concept of the CPF introduced earlier. Suppose that we have a unit of 'resources', which can be used to produce many alternative goods, each yielding as before a set of characteristics. As the number of potential goods gets increasingly large we can think of the CPF in two dimensions as a continuous curve concave to the origin in characteristics space, like the familiar transformation curve in goods space. Which of these infinitely many alternative goods would be the one most preferred by a particular consumer, given his utility function $U(z)$ over the two characteristics? This 'most preferred good' or MPG would obviously be defined by the point of tangency between the CPF and the highest attainable indifference curve, with the slope of a ray from the origin to the optimal point indicating the ratio of the two characteristics provided by the MPG. Other consumers with different tastes would have different MPGs. What should a social planner do if he wants to attain the objective of putting each consumer at a specified utility level with the minimum use of overall resources? In particular, how many and which goods should be produced? With constant returns to scale it is clear that each consumer should be provided with his MPG, in whatever amount is needed to place him at the desired welfare level. With increasing returns to scale, however, we have to trade off the provision of more variety against the sacrifice of less economies of scale. Most of Lancaster (1979) is devoted to a deep and subtle analysis of this fundamental problem under a wide range of alternative technological possibilities, market structures and compensation schemes for the attainment of equitable outcomes, with both first and second-best optima considered. This book, Lancaster's magnum opus, is undoubtedly a major landmark of economic theory that will continue to be an inspiration to the profession for decades to come.

The theory of international trade was another major area that attracted Lancaster's attention and benefited greatly from his application of these novel ideas to it. Early papers on the Heckscher–Ohlin model and the Stolper–Samuelson theorem (see Lancaster 1996, chs 6 and 7) were followed by a pioneering paper (1980) on 'Intra-industry Trade under Perfect Monopolistic Competition', that together with and independently of Paul Krugman (1979, 1980), who was inspired by Dixit and Stiglitz (1977), launched what came to be known as the 'new trade theory', supplementing the standard Ricardian and Heckscher–Ohlin models of perfect competition with models involving economies of scale, differentiated products and monopolistic competition. Unlike the standard models it was easy to show that even identical economies could gain from trade and specialization by providing more variety for consumers in both countries and at lower prices for each differentiated product. The use of the convenient but highly restrictive Dixit–Stiglitz 'love of variety' utility function enabled Krugman to obtain this key result more easily and compactly than the more general framework used by Lancaster; but the latter offers additional insights not available in the former. Later papers considered tariff protection and monopoly policy in open economies in the context of the new trade theory (see Lancaster 1996, Part 1).

At Columbia Lancaster regularly taught in the graduate theory sequence, a course built around his *Mathematical Economics*, an early advanced text published in 1968 the success of which around the world is attested by its translation into Spanish, Japanese, Russian and Rumanian. He also taught a popular undergraduate seminar with the noted philosopher Sidney Morgenbesser. He twice served as chairman of the Economics Department, first from 1973 to 1976 and then from 1989 to 1990. He was elected a Fellow of the Econometric Society, a Distinguished Fellow of the American Economic Association and a Fellow of the American Academy of Arts and Sciences. His death of cancer on 23 July 1999

deprived his university, colleagues, friends and family of a deeply original thinker and a wonderfully warm and compassionate human being. He is survived by his wife Dvora, sons Cliff and Gil, as well as by five grandchildren.

See Also

- ▶ Demand Theory
- ▶ Product Differentiation
- ▶ Second Best
- ▶ Welfare Economics

Selected Works

1956. (With R. G. Lipsey.) The general theory of second best. *Review of Economic Studies* 24, 11–32.
1966. A new approach to consumer theory. *Journal of Political Economy* 74, 132–157.
1968. *Mathematical economics*. New York: Dover, 1987.
1971. *Consumer demand: A new approach*. New York: Columbia University Press.
1975. Socially optimal product differentiation. *American Economic Review* 65, 567–585.
1979. *Variety, equity and efficiency*. New York: Columbia University Press.
1980. Intra-industry trade under perfect monopolistic competition. *Journal of International Economics* 10, 151–175.
1991. *Modern consumer theory*. Aldershot: Edward Elgar.
1996. *Trade, markets and welfare*. Cheltenham: Edward Elgar.

Bibliography

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product variety. *American Economic Review* 67: 297–308.
Krugman, P. 1979. Increasing returns, monopolistic competition and international trade. *Journal of International Economics* 9: 858–864.
Krugman, P. 1980. Scale economies, product differentiation and the pattern of trade. *American Economic Review* 70: 151–175.

Land Markets

Klaus W. Deininger

Abstract

While earlier research highlighted the potential market inefficiencies that can result from the particular characteristics of land, recent empirical evidence suggests that these may be small, not always amenable to policy intervention, and outweighed by the contribution of land markets to broader structural transformations, like population movements out of agriculture. High levels of transaction costs pose, however, still considerable obstacles to land market operation, suggesting that measures to reduce them through greater security and formalization of property rights, a streamlined regulatory framework, and ready availability of information may significantly improve functioning of, and enhance benefits from, land markets.

Keywords

Access to land; Agricultural economics; Banking crises; Fixed-rent contracts; Land markets; Land reform; Land registries; Land tax; Land use rights; Land use regulation; Property rights; Rent control; Repeated games; Sharecropping; Structural change

JEL Classifications

Q15; R14

Land Rental Markets

In a world of perfect information, complete markets and zero transaction costs, the distribution of land ownership will affect welfare but will not matter for efficiency as everyone will operate his or her optimum farm size. However, in most empirical settings, the productivity of land use, and thus the impact of market-mediated transfers

of land, will be affected by technology, producers' ability, potential scale (dis)economies of agricultural production, risk, and imperfections in labour and credit markets. The range of possible contracts will, furthermore, depend on potential tenants' endowments, their reservation utility, and the transaction costs associated with transferring land. Key questions include whether, with a given ownership distribution of land, rental markets will achieve socially desirable outcomes, and which factors will enable participants to attain outcomes closer to the optimum.

By varying the share and a fixed payment to the tenant, landowners who wish to rent can achieve any combination of contractual forms, from a wage labour contract or a share contract to a fixed-rent contract. While all contracts will lead to equivalent outcomes if output is certain and tenants' effort can be enforced (Cheung 1969), relaxation of this assumption gives way to a number of scenarios.

If effort cannot be monitored and agents are risk neutral, only the fixed-rent contract is optimal. The reason is that, in all other cases, equalizing the marginal disutility of effort to their marginal benefit will lead tenants to exert less than the socially optimal amount of effort, thus resulting in lower total production. The optimum outcome will require a trade-off between the risk-reducing properties of the fixed-wage contract, under which the tenant's residual risk is zero, and the incentive effects of the fixed-rent contract, which would result in optimal effort supply but no insurance. Limited tenant wealth has a similar effect because in case of a negative shock tenants with insufficient wealth are likely to default on rent payments. This implies that landlords will tend to enter into fixed-rent contracts only with tenants who are wealthy enough to pay the rent under all possible output realizations, implying that poorer tenants will be offered only a share contract (Shetty 1988). Finally, a dynamic setting opens up a number of additional perspectives, in addition to the scope for using the repeated game context and the threat of eviction to reduce the efficiency losses of sharecropping. A rental contract that provides tenants with adequate incentives to maximize production in any given time period may lead to over-exploitation of the land if (dis)investment is considered, implying that a share contract with lower-powered incentives and possibly compensation may be more appropriate (Ray 2005).

A large literature has focused on testing the extent of inefficiency of sharecropping contracts, although often with mixed results and inappropriate methods (Otsuka and Hayami 1988). Use of within-household variation suggests that, in India, share tenancy is associated with an average loss of productivity of 16 per cent (Shaban 1987) although part of the losses may have been policy-induced. More recent studies fail to find support for inefficiency of sharecropping (Pender and Fafchamps 2006), suggesting that agents' choice of contractual arrangements is rational given the constraints faced in a given situation and that the scope for government to bring about more effective outcomes may be limited.

While potential inefficiencies, if they exist at all, will thus be modest, productivity gains from land rental can be large. Analysis of the same plot before and after being rented in China points towards productivity gains of some 80 per cent, leading to a significant increase in welfare of tenants as well as landlords, in addition to helping the latter to migrate and gain access to non-agricultural income (Deininger and Jin 2006). Although less direct, empirical analysis of determinants for rental market participation in a large number of countries suggests that the ability of those renting in is generally higher than that of those renting out (Deininger 2003), implying a positive productivity impact of land rental which, at least in the case of China, is much superior to what is achieved by a social planner (Deininger and Jin 2005).

The potentially important contribution of land rental to structural change is also illustrated by the fact that rental markets equalize the distribution of per capita operated land area and transfer land to those with lower levels of assets but higher levels of education, and that rental activity increases in settings where wage rates and thus non-agricultural opportunities are higher. Land rental is widespread in developing economies;

71 per cent of farmland is rented in Belgium, and 48 per cent, 47 per cent, and 43 per cent respectively in the Netherlands, France, and the United States (Swinnen and Vranken 2006). Rental markets can emerge rapidly; for example, in Vietnam the share of participants in land rental increased from 3.8 percent of rural households in 1992 to 15.8 per cent in 1998. They were also of great importance in the countries of eastern Europe and the former Soviet Union during the initial phases of economic transition, especially where radical individualization of land was pursued, such as in Albania and Moldova. As long as transaction costs arising from fragmentation were not too high, rental was critical where land had been restored to original owners, many of whom had little intention to use it but also did not want to part with their asset. In West Africa, long-term sharing arrangements did historically provide important incentives for long-term investment and, even though increased population density has shifted contractual parameters in favour of landlords, rental continues to be important in providing land access and increasing productivity.

Of course, a high incidence of rental transactions, and the fact that observed transactions had a positive impact, do not imply that the level of rental activity is optimal. Qualitative and quantitative evidence points towards considerable rationing in rental markets. For example in India, farmers are able to realize only about 75 per cent of their desired level of land transactions (Skoufias 1995), implying that transaction costs or land rental remain high. Two key factors contributing to these are limited security of property rights, which makes renting out too risky, and implicit or explicit restrictions on rental markets in the form of either rent ceilings or the award of property rights to tenants.

Even if it leads to only a small decrease of the probability that landlords who rent out their land will get it back upon termination of the contract, insecurity of property rights can significantly reduce the supply of land to the rental market. This is confirmed by econometric studies in countries as diverse as the Dominican Republic, Nicaragua, China, Ethiopia, Vietnam, and Bulgaria. While insecure tenure may not prevent landlords from renting out completely, it often prompts them to rent only to close kin, where enforcement is easier even if, due to the limited pool of renters to choose from, productivity will be lower than from renting to outsiders, as is indeed observed in the case of Vietnam (Deininger and Jin 2007).

Beyond tenure insecurity, rent ceilings or regulations that aim to confer de facto property rights on tenants by preventing landlords from evicting them and giving heritable use rights to tenants after a certain period of time are a frequent source of inefficiency in rental markets. Although the original intent was to improve equity, such measures led in many cases to self-cultivation by landlords or the adoption of wage labour contracts, both modes of production that are inferior to tenancy in terms of production incentives and outcomes. Analysis shows that, while rent controls can transfer resources to sitting tenants, they tend to make those who are not lucky enough to already sit on tenanted land worse off by restricting the supply of land available to the rental market, undermining tenure security, and reducing investment (Basu and Emerson 2000). Similarly, conferring heritable (but often non-transferable) use rights on tenants, subject to the requirement that they continue paying rent, can increase welfare in the short term but will - in the medium to long term - reduce investment incentives and supply of land to rental markets in a way that is particularly detrimental to the poor and landless, as in the case of India (Deininger et al. 2006) where such legislation has driven a large number of contracts into informality.

Land Sales Markets

Land sales markets provide an opportunity to obtain land for permanent use which will be associated with higher investment incentives than renting. In addition, markets for land sales are a precondition for using land as collateral in credit markets. If all markets were perfect, the sale price of land would equal the net present value of the stream of profits that can be derived from a given land use, and potential buyers would be indifferent between renting land and purchasing

it. However, land sales markets will be affected by a number of factors that include (*a*) the ability to use land as a collateral in credit markets and thus overcome credit constraints; (*b*) expectations about future increases in land values due to infrastructure construction or population growth; (*c*) the risk-return profile and liquidity implications of holding land as compared with other assets; and (*d*) the level of transaction costs in land sales markets.

In economies where risk is high, land is important as a store of wealth, and access to outside credit is limited, land prices can fluctuate significantly over time (Zimmerman and Carter 1999). The reason is that, because returns from agricultural production are highly covariate, demand for land, and therefore land prices, will be high in good crop years when savings are high, sellers are few, and potential buyers of land are many. At the same time, households' need to satisfy basic subsistence needs can give rise to a large supply of land by people who are forced to engage in distress sales of their land in bad years, often to individuals with incomes or assets from outside the local rural economy (Cain 1981). Such distress sales rarely enhance productivity, and improved functioning of markets for insurance and credit to avoid them will be important.

If covariance of asset prices is observed, those who sell off land during crises will not be able to repurchase it during subsequent periods of recovery, creating a potential for successive decline of asset endowments (Zimmerman and Carter 2003). In high-risk environments this may lead the poor to prefer assets with a lower but more stable returns to land even if they had access to credit, implying that, in situations where land is very unequally distributed as in Latin America, land sales markets will not be a good way to achieve asset redistribution, and other measures, such as grants, may be needed to increase land access by the poor on a broader scale.

With macroeconomic instability, an expectation of future land price increases, or lack of sufficiently attractive alternative assets, land may be acquired for speculative rather than productive purposes. For example, inflation and changes in real returns on alternative uses of capital were shown to be key factors explaining changes in land prices in the United States. In eastern European countries, the expectation of large capital inflows due to EU accession was a major factor underlying real estate booms that propelled land prices far beyond the net present value of the flow of services that could be derived from the land. Credit or tax preferences, together with weak regulatory oversight, can reinforce such trends which, in the extreme, can lead to bank crises with far-reaching consequences.

Although empirical study of the functioning of land sales markets is more limited than for rental markets, evidence from India over the 1982–99 period supports the notion that distress sales are important, but that options to insure against risk - for example, the presence of safety net programmes or access to bank branches - helped to reduce or eliminate the adverse impact of climatic shocks. Moreover, there is little evidence of a negative impact of land sales markets on productivity or of speculative land accumulation, partly because of land ownership ceilings and partly because of increased availability of other stores of wealth. Although the number of landless who were able to purchase land remained modest, land sales markets constituted the most important avenue to access land by the poor (Nagarajan et al. 2007).

Well-intended land sales restrictions in a number of countries failed to prevent distress sales but instead drove them into informality. Safety nets and measures to increase access to savings and insurance may be more effective to prevent socially undesirable land loss by the poor. One possible exception is in the transition from customary to more individualized forms of tenure whereby the potential for opportunistic behaviour and land sales by local chiefs is high. To counter this risk, a decision at the local level to maintain a customary land tenure regime that outlaws land transfers outside the community, similar to what was done in the Mexican *ejido* reforms (World Bank 2002), may be an appropriate second-best solution (Andolfatto 2002). As long as it results from a conscious choice and there are transparent mechanisms for changing the tenure regime, such a rule is unlikely to be harmful because, once

potential advantages exceed the cost at the local level, communities are likely to change the rules to allow sales.

Policy Options to Improve the Functioning of Land Markets

Land registries to make information on property rights available publicly in a cost-effective way have many advantages. They reduce the risk of land loss by landlords renting out, and provide the basis for credit market transactions. While informal rights can provide security within a well-defined and socially cohesive group, they preclude trade and exchange beyond this realm. Once gains from transactions with outsiders became sufficiently high, informal rights are likely to be replaced by formalized property right systems and associated enforcement institutions, leading eventually to abstract representation and the impersonal exchange of rights that allows the emergence of more abstract instruments such as mortgages based on the existing rights system (de Soto 2000). Making information on private as well as public land ownership widely available would also reduce the potential for opportunistic behaviour and appropriation of public land by powerful interests as resource values rise.

While disputes among private parties can limit the propensity to rent out land, threats of expropriation without (or with only very limited and delayed) compensation and for a very broadly defined public purpose, which in many countries includes transfer of land to private investors, will limit incentives for investment and can prompt informal pre-emptive land transactions at very low prices that improve neither efficiency nor equity. To prevent this, it is critical to have a restrictive definition of public interest, to ensure compensation at market values if expropriation is unavoidable. If for political reasons ceilings cannot be abandoned altogether, they should be limited to preventing speculative land accumulation. Similarly, land use regulations should be used only if needed to avoid undesirable externalities and if capacity for cost-effective implementation is available.

As public investment in infrastructure and other amenities will be capitalized in land values, taxing land comes close to a benefit tax, and is less distorting than taxes on sales or income. It has thus been considered to be an ideal revenue source for local governments. Land taxes that effectively tax resource rents, that is, that are based on the normal potential yield from a certain plot, will discourage speculation and encourage land owners who are not able to make the most efficient use of their land to rent it out to others. Local land taxes are used effectively in the United States where they have been shown to induce land development. Although underexploited in the past, their potential to intensify land use - which is greater than that of other instruments - has provided a motivation for reforms in a number of countries (Bird 2004).

The often limited ability of the poor to access land through purchase implies that market forces may be unable to correct highly unequal and often inefficient distribution of land, thereby moving the economy towards an equilibrium with a more equal distribution of opportunities and higher overall output. Land reforms in Asia, such as in Japan, Korea, and Taiwan (China), or the abolition of intermediaries in India, and some of the immediate post-independence efforts in Africa - all of which were accomplished under external pressure or immediately after independence - illustrate that land reform can improve household well-being and productive efficiency. At the same time, in many other countries, including virtually all of Latin America, success often remained elusive because, among other things, such measures were guided by short-term political objectives, insufficient effort was devoted to ensuring access to complementary inputs and the competitiveness of producers, and the mechanisms adopted to implement land reform, like ceilings or rent controls, often undermined the functioning of land markets, thus limiting the potential for synergies. Together with multiple restrictions on beneficiaries' ability to transfer the land received, this often limited the scope for land reforms to bring about sustained improvement in beneficiaries' living conditions.

In countries where an unequal distribution of land or incomplete past reforms imply that land reform remains on the agenda, there is broad agreement on a number of common principles (Deininger 2003). These include (*a*) the need to have programmes integrated into a broader development strategy that includes training and capacity building, as well as provisions for complementary investment to make the land productive so as to help put households on a viable trajectory of development; (*b*) a design based on clear and transparent rules that aims to maximize productivity gains; (*c*) a multiplicity of paths to land access needed to underpin land reform, including, in addition to state-sponsored land transfers, progressive land taxation to increase the supply of underutilized land, divestiture of suitable state land, foreclosure of mortgaged land, and rental and sales markets; (*d*) secure and unconditional rights for beneficiaries, including the right to rent or sell their land, perhaps after some initial period; and (*e*) an undistorted policy environment supportive of smallholder agriculture, decentralized implementation, and respect for the rule of law, in particular existing property rights.

See Also

- ▶ Access to Land and Development
- ▶ Agricultural Markets in Developing Countries
- ▶ Common Property Resources
- ▶ Credit Rationing

Bibliography

Andolfatto, D. 2002. A theory of inalienable property rights. *Journal of Political Economy* 110: 382–393.

Basu, K., and P.M. Emerson. 2000. The economics of tenancy rent control. *Economic Journal* 110: 939–962.

Bird, R.M. 2004. *International handbook of land and property taxation*. Cheltenham/Northampton: Edward Elgar.

Cain, M. 1981. Risk and insurance: Perspectives on fertility and agrarian change in India and Bangladesh. *Population and Development Review* 7: 435–474.

Cheung, S.N. 1969. Transaction costs, risk aversion, and the choice of contractual arrangements. *Journal of Law and Economics* 12: 23–42.

de Soto, H. 2000. *The mystery of capital: Why capitalism triumphs in the west and fails everywhere else*. New York: Basic Books.

Deininger, K. 2003. *Land policies for growth and poverty reduction. A world bank policy research report*. New York/ Oxford: World Bank and Oxford University Press.

Deininger, K., and S. Jin. 2005. The potential of land markets in the process of economic development: Evidence from China. *Journal of Development Economics* 78: 241–270.

Deininger, K., and S. Jin. 2006. Productivity and equity effects of rental markets: Evidence from China. Policy Research Working Paper, World Bank.

Deininger, K., and S. Jin. 2007. Does greater tenure security allow more efficiency-enhancing land transactions? Evidence from Vietnam. Policy Research Working Paper, World Bank.

Deininger, K., S. Jin, and H.K. Nagarajan. 2006. Equity and efficiency impacts of rural land market restrictions: Evidence from India. Policy Research Working Paper, World Bank.

Nagarajan, H.K., Deininger, K. and Jin, S. 2007. Market vs. non-market sales transactions in India: Evidence over a 20-year period. Economic and Political Weekly.

Otsuka, K., and Y. Hayami. 1988. Theories of share tenancy: A critical survey. *Economic Development and Cultural Change* 37(1): 31–68.

Pender, J., and M. Fafchamps. 2006. Land lease markets and agricultural efficiency in Ethiopia. *Journal of African Economies* 15: 251–284.

Ray, T. 2005. Sharecropping, land exploitation and land-improving investments. *Japanese Economic Review* 56(2): 127–143.

Shaban, R.A. 1987. Testing between competing models of sharecropping. *Journal of Political Economy* 95: 893–920.

Shetty, S. 1988. Limited liability, wealth differences and tenancy contracts in agrarian economies. *Journal of Development Economics* 29: 1–22.

Skoufias, E. 1995. Household resources, transaction costs, and adjustment through land tenancy. *Land Economics* 71: 42–56.

Swinnen, J.F.M., and L. Vranken. 2006. Patterns of land market development in transition. Policy Research Working Paper, World Bank.

World Bank. 2002. *Mexico - land policy A decade after the Ejido reforms*. Washington, DC: Rural Development and Natural Resources Sector Unit, World Bank.

Zimmerman, F.J., and M.R. Carter. 1999. A dynamic option value for institutional change: Marketable property rights in the Sahel. *American Journal of Agricultural Economics* 81: 467–478.

Zimmerman, F.J., and M.R. Carter. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics* 71: 233–260.

Land Reform

E. V. K. FitzGerald

The redistribution of land property titles by the state is a key issue in poor agrarian countries where land is both the main productive asset and the basis of survival and accumulation for the majority of the population, and thus land tenure is the foundation of the social structure and political power. 'Agrarian reform', which encompasses the transformation of rural administrative institutions, labour use and markets as well, is the modern form of this concept. Urban land reform is not dealt with here, as it is usually subsumed under housing policy. Historically, while widespread changes of land tenure have been characteristic of social revolutions since ancient times (Tuma 1965), and classical economic doctrine supported the sweeping away of the feudal land tenure system to permit commercial modernization and stabilize the independent peasantry; the 'agrarian question' only becomes a central issue of political economy in the 19th century (Hussain and Tribe 1981).

Modern theories of land reform derive from, on the one hand, perceptions of the previous structure of land tenure and production relations; and on the other, the new pattern to be established, intentionally or otherwise. The transition between the two systems can generally be held to involve as central elements both the stabilization of the peasantry and the redefinition of agriculture within the national development model (Ghose 1983).

In capitalist (or 'mixed') economies during the post-World War II years, possibly inspired by the Japanese experience under US occupation, there flourished a considerable enthusiasm for redistributive land reform, which was seen principally as constituting (or reconstituting) a prosperous small-farmer class on estates expropriated from the aristocracy or foreigners (Warriner 1969) with a particular function of underwriting democracy (Jacoby 1971), while responding to the millennarian demands of the peasantry for security (Wolf 1969). However, the international interest in planned economic development led to a concern with the productive consequences of land reform, in terms of both the beneficiaries themselves and urban food supplies; particularly in view of the growing evidence of food output stagnation, underutilized land and rural underemployment (Dorner 1972). The planners' theoretical views on this can be divided (Lehmann 1978) into two groups reaching similar conclusions by different routes. First, there is a structuralist approach (Barraclough 1973; Dorner 1972), stressing the need for more rapid growth of food output to sustain the growing urban wage-bill, underutilization of large estates by 'traditional' landlords, the lack of internal markets for the new infant industries of the import-substitution era, and the necessity to create more rural employment in order to stem migration towards the cities. Second, there is an essentially microeconomic neoclassical approach (Schultz 1964; Griffin 1974; Lipton 1974), emphasizing the superior efficiency of labour-intensive small farmers in terms of land use and the lack of access by peasants to credit and inputs due to tenure structures which permit capital- (or land-) intensive landlord control of markets, which argues that the situation of underemployment of labour and scarcity of capital could be remedied with increased output by redistribution of land titles and the consequent freeing of factor markets.

Such theoretical approaches have had a considerable effect upon the views of international institutions (UN 1976; World Bank 1974) but it would appear that the doctrines applied in practice by governments have been based on objectives more closely related to the maintenance of state power, such as improved supply of cheap food to the towns and the blocking or rural insurgency movements. Land reform projects under these circumstances have involved, at most, the breaking up of large inefficient estates, without affecting commercial farmers (i.e. high land ceilings on owner-cultivated holdings), in favour of individual family farms; with the ultimate purpose of promoting the development of capitalism in agriculture (de Janvry 1981; Ghose 1983). Indeed the so-called 'green revolution' (new crop technologies and mechanization) and resettlement schemes

(moving the landless to areas recently opened up by irrigation or roads) have since the mid-1970s largely replaced land reform in orthodox doctrine (King 1977) as a means of attaining the above objectives without further rural upheaval.

The outcome of such capitalist land reform has been surprisingly similar. Ghose (1983) identifies 'unimodal' pre-reform systems in Asia and the Middle East, where landlords (with merchants and moneylenders) extract economic surplus from small tenant farmers: here the initial effect of land reform is to relieve peasants from this burden, without changing production systems. This raises their incomes substantially but reduces the marketed surplus, while efforts to encourage equitable modernization through the creation of producer cooperatives are undermined by market forces, which tend to lead to peasant differentiation, further encouraged by state support of accumulation by the successful farmer (credits, inputs etc.); eventually polarization between capitalist farmers and landless proletarians results. This transition is more direct in 'bimodal' systems of large commercial estates employing labour from sub-subsistence plots: part of the labour force is 'peasantized' when the land is distributed, but many (particularly migrant labourers) are excluded because there is insufficient land to provide family farms for all, so differentiation starts earlier. A similar polarization occurs in the non-reform sector (de Janvry 1981), the dynamics of which are as important, if not more so, as those of the reformed portion of the land, usually the lesser part in any case. This polarization probably increases productive efficiency, but continued urban food shortages and narrow domestic markets betray the hopes of structuralist theorists; while the slow growth of production and the continued underemployment belie the hopes of the neoclassicals. Except in particular cases, such as Japan, where the state can subsidize the peasant economy out of a highly productive industrial sector; the liberating effect of such land reforms as landlords are eliminated, is outweighed as incipient capitalism eventually disposses the rural poor once more (Ghose 1983).

Land reform is one of the first acts of post-revolutionary socialist regimes (Wadekin 1982), which are faced with the strategic problem of not only collectivizing production relations but also of industrializing a predominantly rural economy while meeting the cost of popular claims for basic needs satisfaction and the defence of the new state against external aggression (Saith 1985). Early socialist thought (Hussain and Tribe 1981) was agreed on the need to sweep away landlordism but not the form of agrarian enterprise that should emerge; indeed it was supposed that capitalism would already have transformed agriculture so that direct worker control along industrial lines would be possible. The experience of Russia and China, however, revealed the need to secure the support of the peasantry, prevent the re-emergence of capitalism, and extract resources (exports, food and labour) from agriculture to finance industrialization. The canonical works of Lenin, Stalin and Mao on this problem have formed the basis for socialist agrarian reform theory, their major differences being in relation to the political role of the peasantry as a 'revolutionary class' (Saith 1985). General agreement on the concepts of land nationalization and eventual collectivization as necessary steps in the construction of socialism meant that doctrine on 'primitive socialist accumulation' made the disposition of the surplus, rather than land tenure as such, the central issue (Saith 1985). This in turn requires a theoretical redefinition of production relations to entail not only the juridical ownership of land as such but also the control over the distribution of its product (Bettelheim 1975).

Land reform doctrine as applied in socialist countries has also revealed a surprising degree of uniformity (Wadekin 1982): in Eastern Europe, as in Soviet Russia four decades earlier, land was nominally nationalized almost immediately but large estates were effectively subdivided among the peasantry, only the more modern ones being retained as state farms; the explicit aim being to secure peasant support for the revolution in the first years. While Lenin had felt that the New Economic Policy could be a vehicle to encourage voluntary cooperativization through favourable internal terms of trade, Stalin implemented forced collectivization culminating in the 'Model Charter' of 1935 appropriating all landed property in the state, establishing collective farms where

income entitlement and certain assets (i.e. a small plot and some livestock) were vested in the household, but which in effect were equivalent to state farms. This model was also applied extensively in Eastern Europe in the early 1950s, although progress was slower in some cases and in others (e.g. Poland and Yugoslavia) the process was not completed at all. However, the only restriction, apart from the avoidance of political destabilization, was to avoid a decrease in agricultural supplies during the transition. The Chinese land reform did not differ in essence from this model as far as tenure is concerned, the operation of nationalized land being vested in the commune with family plots etc.: the major difference was the attention paid to industrial supply to the countryside, and the political emphasis on the transformation of production relations within the collective farm (Lardy 1983).

The more recent attempts to increase rural productivity by various forms of 'liberalization' of socialist agriculture have not involved significant changes in land tenure, but can be termed a 'third agrarian reform' none the less, because they do affect entitlements to the surplus generated on that land, generally in favour of the direct cultivator. In this sense, they can be seen as a 'repeasantization' of agriculture (Saith 1985). At the same time, the experience of the newer socialist states in the Third World has indicated a need to regard export agriculture, rather than food, as the main generator of surplus, because capital equipment and producer goods are mainly imported. These two theoretical advances permit an articulation between various property forms in agriculture where state control is exercised through exchange relations rather than land ownership (FitzGerald 1985).

In sum, we may conclude that modern land reforms 'liberate' the peasantry in their initial stage, strengthening thereby the logic of the peasant economy. Capitalist and socialist land reforms differ in their degree of imposed collectivization and the extent of surplus extraction; but they share the common criterion of planned modernization and thus the ultimate destruction of the peasant economy. Subsequent developments depend upon the national model of accumulation within which agriculture is then inserted. The key factor is not the form of land tenure as such, but rather the use of the economic surplus generated: its retention promotes agrarian capitalism, while its extraction foments peasant resistance.

See Also

- ▶ Agriculture and Economic Development
- ▶ Collective Agriculture
- ▶ Latifundia
- ▶ Peasant Economy
- ▶ Peasants
- ▶ Sharecropping

Bibliography

Barraclough, S. 1973. *Agrarian structure in Latin America*. Lexington: Heath.

Bettelheim, C. 1975. *The transition to socialist economy*. Brighton: Harvester.

de Janvry, A. 1981. *The agrarian question and reformism in Latin America*. Baltimore: Johns Hopkins.

Dorner, P. 1972. *Land reform and economic development*. Harmondsworth: Penguin.

FitzGerald, E.V.K. 1985. The problem of balance in the peripheral socialist economy. *World Development* 13(1): 5–14.

Ghose, A.K. (ed.). 1983. *Agrarian reform in contemporary developing countries*. London: Croom Helm.

Griffin, K. 1974. *The political economy of agrarian change*. London: Macmillan.

Hussain, A., and K. Tribe. 1981. *Marxism and the agrarian question*. London: Macmillan.

Jacoby, E.H. 1971. *Man and the land*. London: Deutsch.

King, R. 1977. *Land reform: A survey*. London: Bell.

Lardy, N. 1983. *Agriculture in China's modern economic development*. Cambridge: Cambridge University Press.

Lehmann, D. 1978. The death of land reform: A polemic. *World Development* 6(3): 339–345.

Lipton, M. 1974. Towards a theory of land reform. In *Agrarian reform and agrarian reformism*, ed. D. Lehmann. London: Faber.

Saith, A. (ed.). 1985. *The agrarian question in socialist transition*. London: Cass.

Schultz, T.W. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.

Tuma, E.H. 1965. *Twenty-six centuries of agrarian reform*. Berkeley: University of California Press.

United Nations. 1976. *Progress in land reform: Sixth report*. New York: United Nations.

Wadekin, K.E. 1982. *Agrarian policies in communist Europe*. Dordrecht: Martinus Nijhoff.

Warriner, D. 1969. *Land reform in principle and practice*. Cambridge: Cambridge University Press.

Wolf, E. 1969. *Peasant wars in the twentieth century*. New York: Harper & Row.
World Bank. 1974. *Land reform*. Washington, DC: International Bank for Reconstruction and Development.

Land Rent

A. Quadrio-Curzio

The history of economic thought can be divided into three broad approaches as to the theory of land rent. Each approach is dominated by a specific concept of rent and tends to prevail in a specific period of the history of economic thought.

The first is the Ricardian approach where, with the premises in the Smithian period and the appendices in the Intermediate period, the surplus theory of rent was laid down. The second is the Marginalist-Neoclassical approach; its forerunner (von Thünen), its most elegant constructor (Wicksteed), its bridge to classical tradition (Marshall) laid down the marginal productivity theory of land rent (and of income distribution). The third is the approach of Sraffa; here, going back to the Ricardian premises, a theory of rent based on an exogenous distributive variable (wage or profit) and on the technical role of 'land' in a modern intersectoral economy is constructed. Let us call it the intersectoral-net product theory of rent.

1. The surplus theory of rent takes origin with the Smithian period: between William Petty's *Treatise* (1662) and Adam Smith's *Wealth of Nations* (1776). The overall vision which comes out of this period is that of rent as a surplus over the cost of production on land (including farmer's income). The size of this surplus depends on the demand of agricultural products and on supply costs which, in turn, also depend on land location and fertility. The receiver of this surplus is the 'class' of landlords.

The core of the surplus theory of rent, however, is given by the Ricardian period, which runs between James Anderson's *Inquiry* (1777) and David Ricardo's *Principles* (1817–1823). The main economists who worked, around Ricardo, on rent in this period were Thomas Robert Malthus, Edward West and Robert Torrens. In 1815 three crucial contributions on rent appeared: Malthus's *The Nature and the Progress of Rent* (which followed *Observation on the Effects of Corn Laws* of 1814); West's *The Application of Capital to Land*; and Ricardo's *Influence of a Low Price of Corn on the Profits of Stock*. The whole discussion found its synthesis in Ricardo's *Principles* (1817–1823), where three famous statements about land rent are made:

(a) Robert Malthus and Edward West 'presented to the world, nearly at the same moment, the true doctrine of rent; without a knowledge of which it is impossible to understand the effect of the progress of wealth on profits and wages' (p. 6); (b) 'rent is that portion of the produce of the earth which is paid to the landlord for the use of the original and indistructible powers of the soil' (p. 67); and (c) 'rent is not a component part of the price of commodities' (p. 78).

Ricardian theory finds the central cause of rent in production (and supply) conditions: rent is due to the growing costs of agricultural production because of decreasing productivity when production is extended. The technical property that land fertility is declining (extensive diminishing returns, which give rise to extensive rent) and that increases in the quantity of labour applied to the same land generates smaller and smaller amount of product (intensive diminishing returns, which give rise to the intensive rent) is the basis of both the laws of diminishing return and of both kinds of rent.

The higher cost (which follows from the technical properties) and price of the last unit of 'corn' produced, in comparison to the previous units of corn, makes rent an unearned surplus for the landlords, the rate of profit being uniform because of competition, and the wage given.

The main corollaries of this principle are:

First, the theory of value: rent is a consequence and not a cause of the high price of corn. Therefore rent does not enter the theory of value. From this point of view, some differences with Smith,

inside the surplus theory of rent, are explicitly pointed out by Ricardo with a clear example: what causes rent is not the demand for 'timber', and its consequently high price; this is in the fact 'the compensation . . . paid for removing and selling the timber, and not for the liberty of growing it' (*Principles*, p. 68). Rent is due only to different costs of production in the use of the productive power of land while profits can increase, ceteris paribus, with an increase of raw materials prices due to a larger demand.

Second, the theory of growth: rent grows with production without technical progress and becomes maximum in the stationary state where profits are zero.

Third, the theory of expenditure: the biggest share of rent is spent on luxury goods. This implies a demand dependence of the luxury goods sectors on the agricultural sector (which in turn affects the other sectors through the wage goods).

2. The marginal productivity theory of rent (and of income distribution) was constructed during the Marginalist or Neoclassical period which runs between 1871 and 1936. Many economists contributed to this broad approach: Jevons, Launhardt, Menger, Wieser, Böhm-Bawerk, Wicksell, J.B. Clark, Hobson, Wicksteed, Marshall, Pareto. However two seem to be most important: Wicksteed and Marshall.

Before considering the theoretical contribution of these two economists we might remember Johan Heinrich von Thünen, who is often considered the forerunner of the new theory. In *Der Isolierte Staat* (1826), he made two specific contributions to rent: on the one hand he constructed a theory of rent on a Ricardian basis but depending upon the most convenient location of various agricultural sectors in relation to demand and market; on the other, he utilized the concept of marginal productivity in a way which could be considered an anticipation of the functional theory of income distribution.

The most concise, elegant and clearcut statement of the marginalist theory of rent was made by Philip Wicksteed in his *Coordination* (1894). We may consider him as representative of this approach.

The new marginal productivity theory of rent (and income distribution) can be considered as a 'consequence' of the Ricardian theory of intensive rent and intensive diminishing returns. Extensive rent and extensive diminishing returns – which was after all the basis of the Ricardian case – are practically neglected because they do not necessarily imply changes in the proportion of the factors of production without which 'there can be neither marginal product nor marginal cost' (Sraffa 1960, p.v). In the extensive case no adaptation to the new theory can be considered satisfactory because of the

> absence of the required kind of change [i.e. in the proportion between factors]. The most familiar case is that of the product of the 'marginal land' in agriculture when land of different qualities are cultivated side by side: on this, one need only refer to Wicksteed, who condemns such a use of the term 'marginal' as a source of 'dire confusion' (Sraffa 1960, pp. v–vi).

Assuming continuous substitutability among factors of production, perfect competition and a special type of production function (linear and homogeneous) rent is determined as the marginal product of land. Furthermore this rule applied to each factor of production (including 'capital') implies the exhaustion of the total product (Euler's theorem). This initial formulation had many successive improvements which are less important to us than the general economic foundation of the new theory.

The main points of the radical change from the classical tradition are: (a) The three-fold division among land, labour and capital is rejected. The unifying element of all 'factors' in the theory of distribution is the service rendered in production. A fundamental symmetry is established: marginal utility of a commodity determines its value; marginal productivity of a factor determines its value. (b) It is useless to consider land rent as an unearned surplus or a residual, it being possible to demonstrate (according to the neoclassical tradition) that the classical surplus theory is 'compatible' with the marginal theory (i.e. the two rents are equal under suitable conditions).

The other economist of the neoclassical tradition to be considered is Alfred Marshall and his *Principles* (1890, 1920); especially because, while trying to stress continuity with classical economists, he did not withdraw from the marginal theory of distribution. Nevertheless he also made more specific contributions. Considering rent as a surplus and a special case of the more general producer's surplus he said:

> . . . rent of land is no unique fact, but simply the chief species of a large genus of economic phenomena; and . . . the theory of the rent of land is no isolated economic doctrine, but merely one of the chief application of a corollary from the general theory of demand and supply; . . . there is a continuous gradation from the true rent of those free gifts which have been appropriated by man, through the income derived from permanent improvements of the soil, to those yielded by farm and factory buildings, steam-engines and less durable goods (Marshall [1890] 1920, p. 522).

Marshall also made notable contributions on the distinction between rent and quasi-rent (problems on which a very important forerunner was Emilio Nazzani 1872), of scarcity rent and differential rent. He stressed the time element in differentiating rent and quasi-rent. As to scarcity rent and differential rent, Marshall concluded that all rents include the two elements.

3. The intersectoral–net product theory of rent is mainly associated with the contribution of Piero Sraffa (1960). The main bases on which Sraffa's theory of rent are founded are:
   (a) A technological-intersectoral scheme: given productive processes which utilize 'land' (a non-produced and scarce means of production) and which produce corn (a basic raw material), the scarcity which 'provides the background from which rent arises' (Sraffa 1960, §88) will appear when (at least) two methods of production are in use for the same commodity. This is the condition for differential rent: the less efficient process utilizing land (zero-rent process) is the basis on which to determine, inside an intersectoral scheme, the prices of all commodities (produced without land), and either the rate of profit (r) or the unit wage (w). Once these are determined, differential rents of the more efficient processes are determined.
   (b) An 'open' theory of income distribution: given r or w exogeneously, rent is determined from technological-intersectoral elements. Furthermore, any change in r or w affects rents via changes in the zero-rent process and/or changes in the cost structures of the rent processes.

On this basis some further points may be considered. The first concerns scarcity, circular processes of production, and the distinction between basic and non-basic commodities. Though rent derives from non-produced means of production and therefore is an element which in a way might be judged as lying outside the core of the Sraffian analytical scheme of circular processes of production, most of its properties can be analysed in that same scheme. The easiest way to understand this apparent inconsistency is to remember that Sraffa assimilates the natural resources (non-produced, scarce, rent-generating) employed in production to the symmetrical category of 'non-basic' commodities which 'although produced, are not used in production' (Sraffa 1960 §85). This symmetry between land and non-basic commodities can be further clarified by considering the effects of taxation: 'Taxes on rent fall wholly on landlords' and 'thus cannot affect the prices of commodities or the rate of profits' (Sraffa 1960, §85). This was also the Ricardian position.

The second point concerns 'fertility' and 'efficiency' and refers to the case of extensive rent which, in the general case, implies m different lands and processes producing corn are in use. Sraffa's theory of rent states that a 'natural' order of fertility of lands does not exist (this means abandoning one of the Ricardian fundamental hypotheses): the set of inputs in the processes cannot be ordered in physical terms. Therefore in the general case, which is the one treated by Sraffa, the degree of 'fertility' depends on the prices of inputs and thus on income distribution. The 'fertility order' is 'not defined independently of the rents; that order, as well as the magnitude of

the rents themselves, may vary with the variation of *r* and *w* (Sraffa 1960, §86).

The third point concerns the case of intensive rent. This is the Ricardian case from which Wicksteed had drawn the marginal productivity theory of rent. Sraffa considers this case as less important than the extensive case; but at the same time he shows its perfect congruence with the criterion that scarcity, which generates rent, implies that more than one method of production with land is in use: 'If land is all of the same quality and it is in short supply, this by itself makes it possible for two different processes or methods of cultivation to be used consistently side by side on similar land determining a uniform rent per acre' (Sraffa 1960, §87). Thus, when an increase of corn production is needed, a second method of production – more productive per acre than the first but with higher cost per unit of product – will be employed on the same land. When the second process has been extended to the whole uniform land, rent disappears; it will reappear when a third process – more costly but more productive – is activated alongside the second in order to satisfy an increased demand of corn. (The intensive rent case, while congruent with the symmetry between land and non-basic commodities implies some more problems in the construction of the standard system.)

In conclusion: the theory of rent developed by Sraffa shows how 'scarcity' can be taken into account in a circular production theory without damaging the foundation of the Classical tradition.

Many extensions followed Sraffa's theory of rent stressing the importance of this contribution and especially going back to the more general problem of the relations between the systems of prices and quantities.

One line of analysis refers to the problem that Sraffa's theory limits the role of land and rent to a situation with fixed quantities. This ignores the effects that a change in the exogenous variable of the price-distribution system (say, *r*) has on the 'efficiency order' of the processes which utilize land, on the number of these processes activated, on the number and the size of rents. And it also ignores the effects that changes in the level of activity produce on the size and the order of the rates of rent, on prices and distribution.

Such issues have been analysed – under the hypothesis of fixed coefficients – by Alberto Quadrio-Curzio (1967, 1980).

The main aspects not determined by Sraffa are:

(1) The distinction of two orders among rents. The first is the 'order of efficiency', which is given by the *signs* (positive or negative) of rents. This order is univocally defined, does not change with the chosen zero-rent process (among the *m* available) and the consequent changes in prices and in the endogenous distributive magnitude. This is the only order to be followed when activating new processes. The second order is that of 'rentability', which is defined among those processes with 'lands' in activity (and therefore all having positive rents). This order is given by the *size* of positive rents and can change when new processes are activated.

(2) The distinction between 'induced' and 'autonomous' changes in income distribution. The induced changes are due to the growing level of activity when *r* (or *w*) remains unchanged. The autonomous changes are due to the change in *r* (or *w*) which can cause changes in the order of efficiency, in the order of rentability, and in the structure and scale of production;

(3) The relative prices of 'corn' always rise, when less efficient 'lands' are utilized, and the relative prices of industrial products ('iron') fall in term of corn.

(4) There is no simple relation between wages and profits, as the role of rents greatly complicates the usual relations between them, also through the choice of techniques and the levels of activity.

(5) The intensive case can be included in the extensive case: in fact, historically, each different land is also the outcome of intensive cultivations. Furthermore, the intensive rent can be considered as a special case of the differential rent: the only condition for

determining the prices of production is that the price of 'corn' covers the cost of production of the less efficient process. This will be zero-rent; whereas the more efficient process will bear a positive rent (for a different treatment of a pure uniform intensive rent case see Montani 1972, 1975).

(6) The dynamic approach is the most obvious development of the intersectoral-net product theory of rent here considered. An approach of this kind, paying special attention to the system of quantities, has been worked out by Quadrio-Curzio (1975, 1986).

Along Sraffian lines other kinds of rents have been pointed out; particularly interesting are the external intensive rent (Abraham-Frois and Berrebi 1980) and the singular rent (Salvadori 1983). Other cases have been considered or can be considered (multiplicity of agricultural products, quasi-rents, exhaustible resources and so on). They show the theoretical possibility of dealing with phenomena of historical relevance in the real dynamics of modern economic systems with the approach based on the intersectoral–net product scheme.

See Also

- ▶ Absolute Rent
- ▶ Corn Model
- ▶ Rent

Bibliography

Abraham-Frois, G., and Berrebi, E. 1980. Rentes, rareté, surprofits. *Economica* 8, Paris.

Anderson, J. 1777. An inquiry into the nature of the Corn-Laws; with a view to the new Corn-Bill proposed for Scotland. Edinburgh.

Cannan, E. 1893–1917. *A history of the theories of production and distribution in English political economy*. 3rd ed. London: P.S. King & Son, 1917.

Gibson, W. 1984. Profit and rent in a classical theory of exhaustible and renewable resources. *Zeitschrift für Nationalökonomie* 44(2): 131–149.

Malthus, T.R. 1814. *Observations on the effects of the corn laws and of a rise or fall in the price of corn on the agriculture and general wealth of the country*. London: J. Johnson.

Malthus, T.R. 1815. *An inquiry into the nature and progress of rent and the principles by which it is regulated*. London: John Murray.

Marshall, A. 1890. *Principles of economics*. 8th ed. London: Macmillan, 1920. Reprinted, 1961.

Montani, G. 1975. Scarce natural resources and income distribution. *Metroeconomica* 27: 68–101. First published in Italian, 1972.

Nazzani, E. 1872. *Sulla rendita fondiaria*. Forli: Tipografia Sociale Democratica.

Petty, W. 1662. A treatise of taxes and contributions. London. In *The economic writings of Sir William Petty*, ed. C.-H. Hull. Cambridge: Cambridge University Press, 1899.

Quadrio-Curzio, A. 1967. *Rendita e distribuzion in un modello economico plurisettoriale*. Milan: Giuffrè.

Quadrio-Curzio, A. 1975. *Accumulazione del capitale e rendita*. Bologna: II Mulino.

Quadrio-Curzio, A. 1980. Rent, income distribution, and orders of efficiency and rentability. In *Essays in the theory of joint production*, ed. L.L. Pasinetti. London: Macmillan. First published in Italian. Bologna: II Mulino, 1975.

Quadrio-Curzio, A. 1986. Technological scarcity: An essay on production and technical change. In *Foundations of economics*, ed. M. Baranzini and R. Scazzieri. Oxford: Basil Blackwell.

Ricardo, D. 1815. An essay on the influence of a low price of corn on the profits of stock; showing the inexpediency of restrictions on importation. London. Reprinted in *The works and correspondence of David Ricardo*, vol. IV, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1817–1823. Principles of political economy and taxation. In *The works and correspondence of David Ricardo*, vol. I, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Salvadori, N. 1983. On a new variety of rent. *Metroeconomica* 35.

Smith, A. An inquiry into the nature and causes of the wealth of nations, ed. E. Cannan. London: Methuen, 1905.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Stigler, G.J. 1941. *Production and distribution theories*. New York: Macmillan.

Thünen, J.H. von. 1826. *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg: F. Perthes. 2nd ed. Rostock: Leopold, 1842–1863.

Valassina, G. 1976. *La teoria della rendita nella storia del pensiero economico*. Milano: Vità e Pensiero.

West, E. 1815. Essay on the application of capital to land, with observations showing the impolicy of any great restriction of the import of corn, and that the Bounty of 1668 did not lower the price of it. London. Reprinted in A reprint of economic tracts, ed. J.H. Hollander. Baltimore: Johns Hopkins Press, 1903.

Wicksteed, P.H. 1894. *An essay on the coordination of the laws of distribution*. London: Macmillan. Reprinted, London: London School of Economics, 1932.

Land Tax

Mason Gaffney

Keywords

Clark, C. G.; Double taxation; Ely, R. T.; Excess burden of taxation; George. H.; Land tax; Neutral taxation; Progressive and regressive taxation; Rent; Stewardship

JEL Classifications

O1

Taxation is the form of socialization used in market economies. Choosing what to tax is choosing what to socialize. Rather than socialize labour or repel capital it is possible to tax land.

Land holds a unique place in the distributional ethic because it is (by definition) of natural origin. Man did not create Earth with its resources but rather fights over it. Land is also (with exceptions) more nearly permanent than man or his works. Thus, rent as private income neither elicits the supply nor preserves it. Its main function is to allocate the fixed supply among uses, but it is arguable that land taxes, when based on land's capacity-to-serve, are at worst neutral to this function and at best improve on it.

The philosophical rationale for land taxes is strongest under an organic theory of the polity. It is no accident that Henry George (prominent protagonist of land taxes) crystallized his ideas after reading Andrew Bisset's *Strength of Nations* on feudal levies. Landholders have a privilege from the state and in return are liable for taxes in perpetuity.

The entire value of land, now and for ever, is here regarded as a benefit received from government. This is consistent with Alfred Marshall's concept, 'the public value of land', where value is the product of three things: nature; government; and spillover values from development of adjoining and linked lands. All these values, being unearned by the individual landholder, are fit to be taxed.

The organic view distinguishes the land from its holder. Land taxes may be paid by income the land earns, not by the holder as a person unless we identify him with the land and regard him as having a prior right to own land free of liabilities to the public from which he holds title. The contractual theory, by contrast, treats government as a kind of business, extending services to specific lands whose holders need pay only for recent benefits received, construed narrowly.

The rationale for land taxes presumes a functional attitude toward distribution, regarding property not as an end in itself but a means to get things done. A land tax based on market value, not varying with actual use, is a fixed cost that sharpens marginal incentives. Critics today seldom argue otherwise, but oppose land taxes precisely because they do force landholders to respond to the market, which may have its own faults in a world of 'second-best'.

Land taxes are *in rem* and so disregard the holder's personal circumstances, a drawback in some opinions. On the other hand landholdings are much more concentrated than the receipt of income or taxable consumption or payrolls, and land taxes are not shifted, making the tax inherently progressive even though but loosely correlated with taxable income. Avoiding land taxes is next to impossible, even though collection enforcement is limited to seizing the land, not the person or any other asset.

The rationale includes a concept of landholder stewardship. A limited number of land titles were issued in order to get land under tenure to assure best use. So far so good, but those not receiving or inheriting land need a counterpoise to assure they receive their share. Land taxes do so in three ways: by supporting government; by pressing landholders to produce goods and services; and pressing them to hire workers to do so. Land taxes act as a kind of social audit and performance standard of stewardship to promote equity towards those excluded.

There is also more equity among landholders, which in turn promotes efficiency. Absent land taxes there is pressure on government to do as much for A's land as for B's. Efficiency, however, calls for specialization and differentiation,

meaning high values for some land and low values for other, with windfalls and wipeouts. Land taxes automatically compensate the losers from the gains of the winners, thus freeing land planners to maximize the joint benefits.

The rationale of equity for the excluded says that lands with open general access like parks and roadways should be exempted in whole or in part. But such exemption can lead to overcrowding, to meet which it is clear that some user charges on such land can be construed as special kinds of land taxes. An obvious example is a charge on large trucks in downtown streets. Lacking any such constraint the crowding might in turn lead to indefinite expansion of the exempt land use.

The rationale is only partly consonant with personal ability to pay. Landholding confers potential ability to pay, but that is only realized upon one's using the land well. And earned cash is not tapped at all. A land tax is a fixed periodic charge. It is based on qualities inherent in the land with few concessions to the landholder's personal illiquidity, weakness, setbacks or ageing. 'Use it or sell it' is the message, which many consider too harsh.

What is harsh for the distressed holder, however, is accommodating to frustrated buyers, and it boils down to which group shall be accommodated. Since liquidity is known not to increase in step with total wealth, imposing taxes on landed but illiquid holders has a strong progressive effect. The regular flow of land taxes also accommodates governments, especially small local ones needing steady revenues that are not turned on and off at the convenience of others.

It is not always a question of selling complete units. Land around homes and enterprises is subject to sharply diminishing marginal utility or productivity and a function of land taxes is to constrain horizontal extension of holdings, to the end that the nucleus of each holding may be closer to others to facilitate trade, cooperation, linkages, sharing common costs, and other synergies. The 'highest and best use' of land is usually that which most relates to and complements its neighbours and trading partners, who must not be held too far distant.

There is also a diminishing return to time as buildings age, and a function of land taxes, in conjunction with building exemption, is to advance (and/or stop retarding) renewal of sites, neighbourhoods, cities, regions and whole economies.

Locke, Quesnay, Adam Smith and others have shown a tendency to shift all taxes to land, whatever the nominal base or event, assuming elastic supplies of labour and capital. This leads some to conclude that all taxes alike just tap land rent. But one cannot tap rent where there is none. Taxes on other bases simply abort the taxed input or activity at the no-rent margins of land use, both extensive and intensive. This excess burden in turn puts an upper limit on the possible tax rate, thus sparing much rent from being taxed at all while destroying other rent completely. The only way to tap much rent is to tax land directly.

Land value and capital are not convertible into one another (excepting exhaustible minerals, not treated here). From this it follows that efficiency does not require equal tax rates on the two, but only uniformity within each class. Uniformity is impossible with capital because of differential concealability. But land is uniformly non-concealable. The case for neutrality of land taxes is stronger under uniformity, but mainly requires that the tax not be a function of use.

A land tax may be based on the current potential rent, or on value. In practice, it is the latter. Values are not simply proportional to rents because many land values are elevated above that by expected higher future rents. In such cases taxes rise high relative to cash flow, and at a stiff rate may even be higher. This subjects the holders to a cash drain. The extra tax may be shown, however, in general to tax the unrealized increment, in the manner advocated by Haig–Simons, at the time it accrues. There is some recent falling-away from Haig–Simons, and to one school now this is 'double taxation', an issue currently mooted.

The most controversial question in land taxation is the effect on appreciating land. Most hands agree the land tax advances conversion to the higher use. To Henry George this 'sovereign remedy' would correct a market failure and unlock speculative holdings with profoundly beneficial effects. To several modern writers following

Richard T. Ely the advance of conversion is unneutral and somewhat wasteful. Speculation is seen as efficiently keeping land from premature commitments. To this writer it seems mathematically obvious that an efficient adaptation to rising future incomes would result in advancing, not retarding conversion. But the issue is now moot.

Land taxation at the local level has a natural cap in local particularism as expressed in 'Don't swamp the lifeboat'. Land taxation by a central national government might go much heavier, and accordingly statesmen like Austen Chamberlain in Britain and James Madison in America have contrived to divert land taxation to local governments. Colin Clark, on the other hand, published a plan to nationalize land through taxation without depriving the poorer localities. He would rank the local jurisdictions in order of land value per capita, and apply a central government surtax starting from zero but graduated upwards according to this ratio. The scheme basically had central government apply to local ones the same principle of direct land taxation that local governments can apply to individuals, tapping the rich rents without destroying marginal rents. Clark, like George, may have been reading Bisset's *Strength of Nations.*

See Also

► George, Henry (1839–1897)

Landry, Adolphe (1874–1956)

E. Malinvaud

Keywords

Böhm-Bawerk, E. von; Fertility; Income distribution theory; Inequality; Landry, A.; Marginal productivity of labour; Mortality; Optimal economic growth; Optimal population size; Production process; Risk aversion; Uncertainty; Unemployment; Value theory

JEL Classifications

B31

Born on 29 September 1874 in Ajaccio, Corsica, Landry died in Paris on 28 August 1956. A graduate of the Ecole Normale Supérieure, he began as a philosopher, then turned to economics and demography. From 1907 on he held the chair of economic history and history of economics at the Ecole Pratique des Hautes Etudes, Paris. He was elected as a Deputy for Corsica in 1910 for the Radical Socialist party, serving as Minister of Navy in 1920, of Public Instruction in 1924, and of Labour in 1932. As a member of Parliament he was particularly effective in promoting family legislation and family allowances which, intended to stimulate fertility, became quite substantial (subsidies to large families in 1913, the Code de la famille in 1939, and the law on family allowances in 1946).

Very early in his study of economics, Landry revealed himself as a gifted theoretician. His approach was purely literary but analytical and rigorous. He was able to master fully technical arguments and, for instance, early exposed in France the definition and relevance of the new demographic indicators proposed by Lotka and Kuczynski. His culture was quite broad and up to date. He was an explicit proponent of the deductive methodology.

His initial concern was with the theory of income distribution. In his dissertation (1901), which made him known as a socialist, he argued that individual ownership and the subsequent unequal distribution of property rights could not be considered as socially optimal and was responsible for a smaller national output than was feasible. His 1904 book was an excellent presentation of the theory of interest in continuation of Böhm-Bawerk, showing why interest was just an aspect of the general theory of value, paying particular attention to the productivity of capital and criticizing Böhm-Bawerk for his overemphasis on the length of the production process. His two articles on the theory of pure profits (1908b and 1938) discussed the role of uncertainty, the idea of risk aversion being already explicit in 1908. However, Landry refrained from making this the first

determinant, arguing that this was rather the scarcity of entrepreneurs, who must simultaneously have capital, abilities and will.

Also interesting are his two long articles. Starting in 1910 from a discussion introduced in 1755 by Cantillon on the demographic impact of a change in landlords' consumption behaviour, Landry finally explains why the returns to primary factors indeed vary with exogenous shifts of individual preferences. Discussing unemployment in 1935, he explains that it reveals an excess of the wage rate over the marginal productivity of labour but is mainly due to a depression of this productivity and can be cured by measures that will raise it again.

From his first writings, Landry always paid attention to population, which later became his main concern. His 1909 article introduced the distinction between three demographic regimes, population being regulated by mortality and the minimum of subsistence in the first case, by fertility behaviour and the wish to achieve some standard of living in the other two, but whereas in the 18th century the objective was a stationary standard of living, it shifted to a permanently progressive one in the late 19th century, 'social capillarity' making this progress feasible for everybody's children. His 1929 article on the optimal size of the population is interesting since it introduces an objective function that was also preferred in the theory of optimal economic growth in the 1960s: a sum of annual terms in which each term is the product of population size and a utility of average consumption per person. His main thesis, developed in his 1934 book, was that a decreasing population leads to decadence, this thesis being substantiated by a study of Ancient Greece and of the cultural centres of the Roman empire.

Selected Works

1901. *L'utilité sociale de la propriété individuelle*. Paris: G. Bellais.
1904. *L'intérêt du capital*. Paris: Giard.
1908a. *Manuel d'économique à l'usage des Facultés de Droit*. Paris: Giard.
1908b. Le problème du profit. *Revue d'économie politique*, January, 24–62.
1909. Les trois théories de la population. *Scientia*. Reprinted in (1934).
1910. Une théorie négligée. De l'influence de la direction de la demande sur la productivité du travail, les salaires, la population. *Revue d'économie politique* 24: 314–328, 364–384, 747–757, 773–785.
1929. Le maximum et l'optimum de population. *Scientia*. Reprinted in (1934).
1934. *La révolution démographique*. Paris: Sirey.
1935–6. Réflexions sur les théories du salaire et du chômage. *Revue d'économie politique*, November–December 1935, 1652–90; March 1936, 327–357.
1938. Sur la théorie du profit. *Revue d'économie politique*, November–December: 1473–1504.
1945. *Traité de démographie*. (In collaboration with H. Bunle, P. Depoid and A. Sauvy.) Paris: Payot.

A fuller bibliography, as well as biographical details appeared in: A. Sauvy, Adolphe Landry. *Population* 11 (1956), 609–620.

L

Lange, Oskar Ryszard (1904–1965)

Tadeusz Kowalik

Abstract

Oskar Lange was worldly known economist, socialist thinker and politician. His special position in economics rested on his profound knowledge of its main currents, of both Marxist economics and Western academic economics (above all the neoclassical) and later of both capitalist and centrally planned eastern European economies. With Abba P. Lerner he was one of the founders of the theory of market socialism. This induced him to make several attempts at a 'major synthesis' and to undertake political actions aiming for a rapprochement between the West and the Communist

world, for peaceful coexistence, economic cooperation and systemic convergence.

Keywords

Accounting prices; Allen, R. D. G.; Breit, M.; Brus, W.; Business cycle; Cash balances, demand for; Command economy; Concentration; Corporations; Democracy; Depreciation; Econometrics; Economic calculation in socialist economies; Efficient allocation; Forecasting; General equilibrium; Hicks, J. R.; Industrialization; Innovation; Interest, theory of; Interpersonal utility comparisons; Inventories; Kalecki, M.; Keynesian Revolution; Knight, F. H.; Lange, O. R.; Lange–Lerner mechanism; Laski, K.; Leontief, W.; Lerner, A. P.; Liquidity preference; Loanable funds; Marginal analysis; Marginal cost pricing; Marginal efficiency of capital; Market socialism; Marx, K. H.; Miller, D.; Mixed economy; Money; Money supply; Monopoly capitalism; Neutrality of money; Optimal resource allocation; Planning; Price flexibility; Propensity to consume; Public ownership; Public trusts; Roemer, J. E.; Samuelson, P. A.; Say's Law; Schultz, H.; Schumpeter, J. A.; Stiglitz, J.; Sweezy, P. M.; Wage differentials; Yunker, J. A.

JEL Classifications

B31

Lange was born on 27 July 1904 in Tomaszow Mazowiecki, near Lodz, Poland, into the family of a German-born, assimilated textile manufacturer, and died on 2 October 1965 in a London hospital following thigh surgery. He studied law and economics in Poznan and Cracow. His main tutor was Adam Krzyzanowski, liberal and Anglophile. In 1929, Lange studied in London and in 1934–5 in the United States, mostly at Harvard and Berkeley. He lectured in statistics and economics in Cracow (1927–37), Chicago (1938–45) and Warsaw (1948–65). Politically involved since his youth, he was active at the Independent Socialist Youth Union in the interwar period.

During the Second World War he pushed the cause of Soviet–American rapprochement and socialist–communist cooperation. He served as the first ambassador of the Polish People's Republic in Washington (1945–6) and as the Polish delegate to the UN Security Council (1946–7). Later he was a member of parliament and a member of the State Council in Poland.

Lange's special position in economic theory rested on his profound knowledge of its main currents, of both Marxist economics and Western academic economics (above all the neoclassical) and later of both capitalist and centrally planned Eastern European socialist economies. This induced him to make several attempts at a 'major synthesis' and to undertake political actions for a rapprochement between the West and the Communist world, for peaceful coexistence and economic cooperation.

Capitalism and Economics

The capitalist economy was Lange's chief research concern from his early youth until the end of the Second World War. His primary interests included the study of business cycles and the evolution of capitalism. His Ph.D. thesis was a study of business cycles in the Polish economy 1923–7 (1928a), and won the title of docent (assistant professor) for a statistical study of the business cycle (1931a). These were among the chief topics of his lectures at US universities, mainly in Chicago. Early in the war he studied, together with L. Hurwicz, ways of empirical verification of business cycle theories. Although he became a leading authority on this subject (see his review, 1941a, of Schumpeter's book and, 1941b, on Kalecki's cycle theory), he never produced a complete theory of his own. His studies of the business cycle led him to econometrics, a discipline he helped create (during the Second World War he edited the quarterly *Econometrica).* His textbook of econometrics (1959), the first of its kind in eastern European countries, recapitulates his studies of business cycle and of market mechanisms, in addition to an outline of programming

theory based on Leontief's input–output tables and on Marxian reproduction schemata.

The evolution of capitalism was a close interest both as a scholar and as a political writer. Initially he believed that the development of large corporations marked a transition from 'the anarchical freemarket capitalist economy to a consciously planned economy' (1929 [1973, p. 70]), that is, to an organized capitalism. But with the Great Depression those hopes vanished. Monopolies and government intervention cause chaos and disarray in the economy and led eventually to a collapse of capitalism and the victory of socialism (1931b [1973]). Soon, however, he came to the conclusion that 'it was not capitalism but the worker movement which collapsed during the crisis' resulting in a 'stabilization of capitalism' (1933 [1973, p. 63]).

Just before and during the war, Lange often argued that capitalism cannot possibly be reconciled with economic progress in the long run. But at the same time he looked for ways of reforming capitalist structures to turn them into mixed-type economies –calling for a socialization of the monopolies which he regarded as threats to political democracy and which he blamed for generating unemployment.

During his stay in the United States, Lange published a number of contributions exploring and developing, as well as criticizing, the standard economics which was, and continues to be, taught at most universities in the West. Those studies fall roughly into two categories: the first was 'pre-Keynesian' from the point of view of general approach, while the other was closely connected with the absorption of the 'Keynesian Revolution' by traditional economics.

In one major study (1936b, 1937b), Lange tried to explore the relationships between interest theory and the theory of production factor cost. Using a strongly simplified model (one final commodity produced by labour and one capital good, free competition, 'neutral' role of money, risk is neglected), Lange unfolded a theory of interest which in many of its points came close to that of Frank Knight, even though in his concept of money capital ('as a general command over means of production') he was influenced more strongly by Schumpeter and Marx.

Lange is regarded as one of the founders of 'modern welfare economics' (Graaff 1957). Following Bergson's pioneering study (Burk 1938), Lange listed (1942a) theorems, which do not require interpersonal comparability of utility as well as those which do. The study of optimal distribution of incomes must be based on a priori hypotheses concerning marginal utility of incomes for different persons. For welfare economics propositions it is not necessary that utilities of individuals must be measurable as long as these utilities can be ordered.

The next and probably most important group of studies concern Keynesian theory's relationship to the mainstream of Western economic thinking. In a (1938b) study, Lange explores the internal logic of Keynes's theory investigating the mutual relations between interest rate, propensity to consume, marginal efficiency of capital, investment and national income. In Lange's model, elasticity is the all-decisive concept. Using this concept and some of Walras's ideas, Lange outlined a 'general theory' of which the Keynesian theory was one particular case. That special case occurs when elasticity of liquidity preference to income is close to zero or when it is infinitely great in relation to the rate of interest. Then, the rate of interest does not depend on marginal efficiency of capital or on propensity to consume. When the elasticity of liquidity preference to the rate of interest is close to zero, then the classical and neoclassical theory, stressing the dependence of money demand on income alone, holds. Keynes approved Lange's interpretation of his theory as following 'closely and accurately my line of thought' (Keynes 1973, p. 232n). Lange's exposition of the notion of multiplier (1943a) was more modest in its intention.

Analysing Say's Law (1942b), Lange made one of the first ever attempts to overcome what was called the dichotomy of the pricing process. In traditional neoclassical theory, commodity prices were determined under the assumption that money is just 'a worthless medium of exchange and a standard of value' (1942b,

p. 64), and hence of a barter economy. Only later on, prices determined in this way, were pecuniary prices 'superimposed'. Accordingly, the substitution of money for commodities and vice versa was ignored completely. That was the gist of the assumption that total demand is identically equal to the total supply of commodities. Thus, the theory of money must start with the rejection of this contention (of Say's Law) and investigate conditions and processes leading to equilibrium of total demand with total supply. For this purpose, money must be included in the theory of general equilibrium.

These studies prepared the ground for a more ambitious synthesis. In his previous studies, Lange had already studied questions and problems asked by Keynes (this partly holds also for the theorists of imperfect competition and for Schumpeter) and tried to resolve them in his own fashion, relying on mathematical tools of general economic equilibrium as developed and modified by Henry Schultz, R.G.D. Allen and Paul Samuelson, but especially by J.R. Hicks.

That undertaking found its most complete and systematic exposition in Lange's (1944a) book, which sums up his theoretical work during his American period. The book is something like a restatement of the theory of general economic equilibrium in which money is incorporated explicitly as part of this theory. Substitution between money and goods is the key concept for understanding processes of equilibrating and disequilibrating the national economy. As Lange puts it, 'The interest in the problem and the recognition of the crucial importance of substitution between money and goods were inspired by Lord Keynes. For the tools of analysis the author is heavily indebted to Professor J.R. Hicks' (1944a, p. vii).

But Lange's book was an outcome as much of theoretical as of practical disputes over general economic policy. His main point of interest was the belief, which survived repeated attacks from Keynesians, that price flexibility – and in particular flexible prices of production factors, mainly of labour – is a condition of full utilization of production factors. Defending the Keynesians' position on this matter, Lange intended to reach both the general public and sophisticated, mathematically minded economists who refuted Keynes's language of aggregate concepts as too unscientific.

With a view of such different audiences, Lange composed his exposition at two or even three levels of difficulty. The main body of the book is 'as simple as possible' and in colloquial non-mathematical language full of socio-political corollaries. Only in the numerous footnotes did he present technical details. The final part of the book, called 'The Stability of Economic Equilibrium' and published as an appendix, is in rigid mathematical language and is addressed to the narrower group of specialists.

The book's main message can be summarized in the following way. There are three ways in which money can affect economic equilibrium under flexible prices:

1. If the overall amount of money is constant, the fall in prices of a factor leads at first to a fall in other prices and to a growth in purchasing power of the existing stock of money. An excess supply of money arises. This, in turn, drives up demand for goods and checks prices from falling further. As other prices are falling less quickly than that of the factor under consideration, demand for this factor increases. Along with that, the amount of loanable funds grows, which causes a fall of the interest rate. This, then, encourages investment and results in employment growth. This is the case of the effect of money being *positive.*
2. When the overall amount of money is determined by credit creation and changes in step with the changing demand for money (cash balances), the effect of money can be said to be *neutral.* In this case, the mechanism of automatic maintenance and restoration of equilibrium no longer works. The stock of money shrinks in proportion to the falling demand for cash balances and an excess money supply develops. The purchasing power of the stock of money remains unchanged. In consequence, the fall in prices is not checked by a rise in the purchasing power of the stock of money and interest rates do not fall. The excess supply of

the production factor under consideration is not being absorbed.

3. Money has a negative effect when its amount shrinks more than proportionately to falling demand for cash balances. Banks, for example, react to the fall in prices by demanding loan repayment. A shortage of money is then felt in the market. Pessimism, growing uncertainty, and so on fosters this development. Then, a fall in the given production factor's price (for example, wages) causes an even more dramatic fall in prices of other goods, which leads to an even larger excess supply of the production factor than was the case originally (for example, to even higher unemployment).

Lange's general conclusion from his analysis was quite pessimistic:

> Only under very special conditions does price flexibility result in the automatic maintenance of restoration of equilibrium of demand for and supply of factors of production. These conditions require the combination of such a responsiveness of the monetary system and such elasticities of price expectations as produce a positive monetary effect, sensitivity of intertemporal substitution to changes in interest rates ..., absence of highly specialized factors with demand or supply dependent on strongly elastic price expectations, and finally, absence of oligopolistic or oligopsonistic rigidities of output and input. To a certain extent, the absence of a positive monetary effect may be replaced by the stabilizing influence of foreign trade ... (1944a, p. 83)

On the whole, Lange regarded price flexibility as 'a workable norm' of long-run but not necessarily short-run economic policy during the long period of between the 1840s and 1914. However, the favourable conditions which prevailed during that period belong to the remote past. The oligopolization process, the deteriorating investment opportunities, the tendency towards money supply caused by new technology applications, along with the bad experiences of the two world wars and the Great Depression – all these made any automatic attainment of equilibrium and stability a very unlikely prospect.

This conclusion prompts two questions. First, what significance does the general economic equilibrium theory have for economic theory and for economic policy? Several years later, Lange compared that theory, which deals with very unlikely contingencies, to the case of an ape trying to write the Encyclopaedia Britannica. While probability calculus does not preclude such a possibility, we should ask ourselves if dealing with such an unlikely case is not an utterly futile exercise.

Price flexibility was the last fruit of Lange's study of the general equilibrium theory. To what extent his subsequent silence on this subject was due to the fact that, after 1945, he found himself in an entirely different environment, and to what extent due to his disenchantment with the theory, is difficult to say. Anyway, his economic thinking in later years took an unexpected turn. Contrary to his attitude in public life, as a philosopher of science Lange was rather conservative-minded, believing that 'science does not progress ... by the wholesale rejection of old theories and the devising of new ones, but by arduous work of enriching and improving existing scientific achievements' (1970, pp. 80–1). Accordingly, he put a great deal of effort into showing that the so-called Keynesian Revolution was no revolution at all; and that it should be viewed as a contribution merely 'enriching and improving scientific achievements'. But when he accomplished that job, Lange dropped the synthesis he had worked out with such a great expense of effort only to choose an alternative paradigm.

After the Second World War, however, Lange only sporadically resumed his study of capitalism, mainly to consider whether capitalism is able to resolve economic problems of backward countries (to which his answer was emphatically negative, 1957) or prospects for disarmament and economic cooperation between the Council for Mutual Economic Assistance countries and the capitalist West.

Lange–Breit Model of Socialist Economy

Lange first manifested himself as a socialist writer in his book (1928b) on Edward Abramowski (1868–1918), whose ideology Lange called 'constructive anarchism'. In those ideas, Lange emphasized Abramowski's resentment of

government interventionism, pitting it against the ideas of English Guild Socialism and of Austro-Marxism, both of which had strongly influenced Lange himself. Lange advocated especially the idea of industrial self-government, of separating the economy from political power, and the decay of the state as an institution of class domination though not of an instrument of coercion.

Together with MarekBreit (1907–42), he wrote the first outline of a socialist economy's functioning in the chapter of a collective book, *Economy–Polity–Tactics–Organization of Socialism* (1934 [1973]). It was the product of a group of left-wing socialists, led by Lange, and committed to the revolutionary reconstruction of a system in Poland, which would be different from the Soviet model of polity and economy.

The Lange–Breitmodel, or the 1934 model (see Kowalik 1970, 1974; Chilosi 1986, 2005; Toporowski 2003) is one version of a corporate market economy under socialism. It rests on the following rules. Plants should go public, or be 'socialized', in his terminology, by transferring private ownership titles to a Public Bank and by organizing the national economy into public trusts by industrial branches. Trusts would be the basic units of the economy and endowed with a great deal of autonomy. The decisive say in their boards would belong to workers, who would be organized into 'an appropriate system of worker councils'. Trusts autonomy is limited by the Public Bank's supervision and coordination functions or, more exactly, by the functions performed by a uniform and monopolistic bank system. Basic planning instruments would include accumulation fund management and trust financing. The Public Bank would also watch if trusts and companies subordinate to them abided by management rules, in particular by rules of 'rigorous' price and cost accounting. Plants run at a loss would be closed down. Plants failing to record an average surplus would forfeit their right to get loans not only for expansion but even for ordinary capital replacement, and hence they would decline. Both trusts and plants would be obliged not only to remit their production costs but also to achieve a certain accumulation, the rate of which would be established by the Public Bank and subsequently redistributed for investment and for subsidizing public utilities (which may be run at a loss).

Since trusts would hold virtually monopoly power in the market, as all public plants would by law belong to some trust, Lange and Breit perceived the danger of charging excessive prices and cutting output rates. They realized that such a policy might become quite popular among employees of any given trust, who might hope to get their wages increased. To forestall monopoly practices, they therefore proposed to oblige trusts to take on all job-seekers applying to them. If price increases resulted in higher wages in any given trust, employees from other trusts would swarm to it so that the increased wage fund would have to be redistributed among a larger number of employees. The underlying purpose of that obligation, then, was to deter trusts from driving up prices.

As the two authors did not consider the question of inflation, they did not say why excessive wage increases by one trust should not set off an avalanche of price increases if other trusts attempted to forestall an exodus of their own workforce. Nor did they envisage possible consequences of the indivisible nature of means of production and of possible consequences of delays in market adaptation. Moreover, the Public Bank's investment policy would be based on workforce migration in reaction to changing demand, price fluctuations and subsequently price changes. This was to be something like an automatic indicator of demand intensity for individual goods.

The Public Bank would further control capital imports and exports, whereas a 'foreign trade office' created by the trusts concerned would be in charge of goods sales and purchases abroad. The Public Bank would also be authorized to transfer capital assets from trust to trust.

The private sector, which is consistently referred to as the 'non-socialized', that is, non-public, sector of the economy, was to remain 'broad', consisting of private farms holding less than 20 hectares of land, crafts shops, business enterprises with less than 20 people on their payrolls, as well as retail trade shops. However, because economies of scale were expected to

impart higher efficiency to larger companies, the private sector would be 'a relic on the way out'. The two authors said nothing about credit policies towards this sector, but the Public Bank would conduct a discriminatory kind of policy towards profit-making small capitalist businesses (up to 20 employees) designed eventually to bring about their demise through taxes. Lange and Breit recommended that the Public Bank should levy taxes equal to the accumulation rate, which was supposed to reduce owners' incomes to the level of manager's salaries. The two authors failed to take account of the role of risk and innovation.

Nor is it clear how the two authors thought plants (which they preferred not to call enterprises) would be managed, or how trusts would be organized and what prerogatives the latter would have. They merely said workers organized in a system of worker councils would have the decisive say and that trade unions and worker cooperatives were best suited to create trusts. Nor did they propose any clear procedure for appointing the Public Bank's board of management, which was expected to make the socialist economy a planned economy.

Designed as an alternative model to the command–planning system then existing in the Soviet Union, the Lange–Breit concept was largely reminiscent of Bolshevik concepts from before the period of wartime communism or right after it (trusts, worker councils, a single state-owned bank, a long-run policy of farm collectivization), modified by an emphasis on separating political authority from economic organization, on impartial economic criteria, and on recognizing consumer preferences as the foundation of investment policies.

The Theory of Market Socialism

The next model of socialist economy, which I propose to call the classical, Lange presented in a study (originally published as two articles, 1936a, 1937a), and in a book form (with Taylor 1938b). It was devised only two or three years after publishing Lange–Breit model. But this period brought an immense improvement of Lange's analytical expertise.

On a Rockefeller Foundation Grant, Lange studied at Harvard, Berkeley and Chicago, and at the London School of Economics. He was strongly influenced by Schumpeter, under whose tutorship he worked at Harvard during most of his two- year scholarship, and he took part in a famous seminar (The Economics Club) led by the Austrian-American economist. That influence surfaces in many of Lange's studies, including his study *On the Economic Theory of Socialism*, especially in the economic justification of socialism. That study, or at least its main body, was written at Harvard and must have been heatedly discussed there. At that time he also became intellectually involved with the brothers Alan and Paul Sweezy, economists and socialists of a similar orientation to that of the visitor from Poland. He also had working contact with W. Leontief.

On the Economic Theory of Socialism expresses Lange's long-lasting conviction that neoclassical economics, especially welfare economics, is best suited to serve as a foundation of a theory of socialist economy.

The classical model, of course, is theoretically more sophisticated and more accurate in its purely economic aspect, but perhaps at the cost of giving less specific treatment to institutional aspects than the 1934 model. That was probably due to the chief purpose of that study, namely, to disprove Mises' argument about a theoretical and practical (practical, according to Hayek and Robbins) infeasibility of economic calculus in socialism because of the absence of a genuine market (prices) for capital.

Many formulations in that classical study indicate that a socialist society's general outlines of economic organization were similar or identical in both the early and classical models. In particular, this is true of the separation of political power from economic management, of its three-level structure – the centre, the branches organized in trusts, individual plants – and of the similar powers of the Central Planning Board (CPB) and the Public Bank. In both models, the centre is expected to react to changes in market factors (prices and wages) and, correspondingly, to changes in employment in the early model or to changing inventories and emerging shortages in

the classical one. The CPB, basically, is to imitate the market. The early model was clearly more 'market-oriented' because all prices of goods and services were to be determined by the market. Accordingly, there would be no difference between actual market prices and calculated prices as set by the CPB.

Lange–Lerner Mechanism

This is a designation commonly used to denote a market-oriented socialism model devised by Lange, who later amended it after public discussion with Lerner. The first, fundamental part of Lange's study was published together with A.P. Lerner's (1936) critical remarks in the same issue of the *Review of Economic Studies,* while the second part appeared together with Lange's reply to Lerner (1937). Later on, Lange made the changes necessary to publish his study (together with F.M. Taylor's essay) in book form (1938b). The term is occasionally used in a less restricted sense, to bring out the similarity of Lange's and Lerner's views on other matters concerning socialist economy.

The mechanism of socialist economy in the Lange–Lerner blueprint was based on the following assumptions. It has its institutional framework in the public ownership of means of production (for simplicity, the private sector is omitted) and in the free choice of consumption and employment (job and workplace), while consumer preferences –'through demand prices'– are the all-decisive criterion of both production and resource allocation. Under these assumptions, an authentic market (in the institutional sense) exists for consumer goods and labour services. But prices of capital goods and 'all other productive resources except labour' are set by a CPB as indicators of existing alternatives established for the purpose of economic calculation. So, apart from market prices, there are also 'accounting prices'. In order to make their choices, both categories of prices are used by enterprise and industry managers, who are public officials.

Production managers in charge of individual enterprises or entire industries make autonomous decisions about what and how much should be produced and how it should be done, while prices are set as parameters outside the enterprises or industries. But since profit maximization has by definition ceased to be a direct goal of economic activity, to ensure that they can achieve effects close to those achieved in free-market economy, production managers must obey two rules. First, they must pick a combination of production factors under which average cost is minimized; and second, they must determine a given industry's total output at a level at which marginal cost is equal to product price. The first rule was expected to eliminate all less efficient alternatives. In combination with the second rule, in so far as it concerns plant managers, it performs the same function as the free-market economy desire to maximize profit. This leads to minimization of production costs. The second rule compels production managers to increase or cut the output of a whole industry in accordance with consumer preferences, which is a substitute for free entry in a free competitive economy.

These rules lead to an economic equilibrium by the trial-and-error method first described by Fred M. Taylor (1929). The CPB acts like an auctioneer, initially watching the behaviour of economic actors in reaction to a price system it picks at random or – perhaps the best solution – to the historically inherited prices. The behaviour of the system is measured by the movement of inventories of goods. If there is too much of some product at a given price, then its inventory grows, and vice versa. This is regarded as information that the product price should be cut or increased, respectively. This procedure is applied as many times as is necessary to reach equilibrium, providing that this process does in fact converge to the system of equilibrium prices. Accounting prices, then, are objective in character, just like market prices in a competitive system, the difference being that in this case the CPB performs the role of the market.

The same trial-and-error way towards equilibrium could also be applied in two other models of socialist economy, one providing for a decreased consumer influence on production programme, the other presupposing none at all.

In its extreme version, which for sociopolitical reasons Lange deems untenable, the model might provide no freedom of choice for either consumption or employment. Production plans would be decided by the CPB officials' scale of preferences. In such a version all prices are basically accounting prices. Consumer goods are rationed, while the place and kind of employment are imposed by command. If production managers keep to the above-mentioned rules, and if the CPB keeps to the parametric price system, then economic calculus is possible even in this version, while prices are not arbitrary but reflect the relative scarcity of factors of production.

There is an intermediate model, which provides for freedom of consumption decisions but only within a production plan established on the ground of CPB preferences. In this case, accounting prices of producer and consumer goods reflect the CPB's preference scale, while production managers would rely on them in their decision-making. Market prices for consumer goods would be set by supply and demand. But Lange rejects even this system as undemocratic, saying that the dual system of prices could be applied only when there is widespread agreement that checking the consumption of some products (say, alcohol) while promoting the consumption of other goods (say, cultural services) is in the public interest.

But the CPB might conceal its preferences and resort to rationing production goods and resources. Society can defend itself against such practices by creating a supreme economic court, which would be entitled to declare any unconstitutional CPB decision as null and void. In Lange's view, any decision introducing rationing would be unconstitutional.

Interestingly, Lange rejects these two versions of socialist economy on account of the potential hazards they carry for democracy, and says not a word about democracy's possible link with economic efficiency.

Lange considers the distribution of national income in three aspects.

Wages would be differentiated by seeking a distribution of labour services that would maximize society's wealth in general. This happens when differences in marginal disutility of work in different trades and workplaces are offset by wage differences. Wage differentials can be treated as converses of prices paid by employees for differing work conditions, as a simplified form of buying free time, safety or pleasant work (which is easy to imagine assuming that all employees get the same earnings but pay different prices for doing different jobs; the easier and safer a given job, the more one has to pay for it). In this sense, the wage differentiation rule can be brought into harmony with egalitarianism.

Apart from wages paid by employees, each consumer is paid a public dividend as his or her share of capital and natural resources. At first Lange was inclined to distribute such dividends proportionally to wages. But as Lerner pointed out that such a policy would impart added attractiveness to the hardest jobs, Lange changed his mind, saying there should be no link between procedures for public dividend distribution and wage differentials.

The distribution of national income between consumption and accumulation, said Lange, would not be arbitrary when only consumers' individual savings decide the rate of accumulation. But if savings are 'corporately' determined – and Lange at first thought that was typical of a socialist economy – then there would be no way of preventing the CPB from being at least partly arbitrary in its decisions.

Emphasizing that resource allocation is guided by formally analogous rules in both socialist and free competitive economies, Lange argued that real allocation in socialism would be different from and more rational than that in capitalism. In his static analysis, he considered the following factors as decisive in judging the relative performance of the two systems. Greater equality of income distribution enhances society's well-being (in the subjective sense, that is, as a sum total of individual satisfactions). Second, socialist economy makes allowances in its calculus for all the services rendered by producers and for all the costs involved, while a private entrepreneur does not care for benefits that do not flow into his own pocket nor for costs he does not have to pay: 'Most important alternatives, like life, security, and health of the workers, are sacrificed without

being accounted for as a cost of production' (1938b, p. 104).

Even the possible flaws that Lange conceded might appear in a socialist economy, such as the arbitrary setting of the rate of accumulation or the danger of bureaucratization of economic life, would be milder than under capitalism, he argued.

But the ultimately decisive economic argument in favour of socialism, Lange believed, was the general waste and endogenous tendency towards stagnation generated by modern capitalism's monopolistic tendencies. This question, though, goes beyond the scope of the often-criticized static analysis underlying Lange's classical model. Leaving aside the now enormous critical literature, let us try to answer the question of what Lange himself saw as his model's limitations.

Lange anticipated possible charges by critics in the second part of his study, in his discussion of 'The Economist's Case for Socialism':

> The really important point in discussing the economic merits of socialism is not that of comparing the equilibrium position of a socialist and of a capitalist economy with respect to social welfare. Interesting as such a comparison is for the economic theorist, it is not the real issue in the discussion of socialism. The real issue is *whether the further maintenance of the capitalist system is compatible with economic progress.* (1938b, p. 110)

But as he develops this general idea, Lange clearly uses an asymmetrical kind of argument. Having presented free competitive capitalism as the system that generated 'the greatest economic progress in human history', Lange proceeds to show (among other things, by referring to Keynes) that the source of that progress is drying up because of the progressive concentration and monopolization of production. His main point is that corporations, which are capable of controlling the market, attempt to avoid losses due to capital depreciation caused by innovation, and hence they try to check progress in technology. Neither a return to free competition nor government control can effectively eliminate this tendency. The only effective solution, then, is the socialization of big capital, the introduction of socialism.

But will socialism ensure rapid technical progress? Will the abolition, via socialization, of capitalist monopolies' well-known tendency to check technological progress automatically dismantle all the barriers to innovation? Or will it amount to substituting new barriers for old? Will the two rules for managers be sufficient to guarantee the adoption of state-of-the-art production techniques? In his classic study, Lange never even asked such questions and only much later did he become aware of them.

Towards the end of his life (in a letter to the present writer dated 14 August 1964), Lange wrote:

> What is called optimal allocation is a second-rate matter, what is really of prime importance is that of incentives for the growth of productive forces (accumulation and progress in technology). This is the true meaning of, so to say, 'rationality'.

It seems that he must have lacked the indispensable tools to solve this question or even to present it in detail.

Towards a Mixed Economy

Perhaps, the most important difference between the early (Lange–Breit) and the classical models was his new emphasis that 'the real danger of socialism is that of a bureaucratization of economic life, and not the impossibility of coping with the problem of allocation of resources'. He reassured himself by pointing out that the same danger existed in monopolistic capitalism and that 'officials subject to democratic control seem preferable to private corporation executives who practically are responsible to nobody' (Lange 1938b, pp. 127–8).

When he became aware of that danger, which would exist even in a market- dominated brand of socialism, he embarked on a long quest for what he called in the title of one article (1943b), 'The Economic Foundations of Democracy in Poland'. In the classical study he had already put forward the idea of a Supreme Economic Court whose function would be to safeguard the use of the nation's productive resources in accordance with the public interest, in particular to declare as null and void any CPB decision which was incompatible with adopted management rules.

During the Second World War, Lange suggested a number of ideas for better safeguards for

democracy, either by substantiating the injunction to take account of consumer preferences (and hence limiting the central economic authority's prerogatives) or by devising institutional guarantees for democratic control of decision-making bodies, or by indicating limits to the socialization of property.

There were a number of highlights of the evolution of Lange's views during that period.

In his letter to Hayek in 1940 (Kowalik 1984) Lange gave a more accurate, and perhaps slightly different, description of the CPB's prerogatives for pricing goods and services:

> Practically, I should, of course, recommend the determination of prices by a thorough market process whenever this is feasible, i.e. whenever the number of selling and purchasing units is sufficiently large. Only where the number of these units is so small that a situation of oligopoly, oligopsony, or bilateral monopoly would obtain, would I advocate price fixing by public agency . . ..

Accordingly, he recommends socialization of industries only in areas where there is not automatic competitive market process.

Later in 1942–3, he departed even further from his classical model towards a mixed economy. In his review of Dickinson's book (1942c), he had the following idea of how to prevent the central authority's arbitrariness in determining the accumulation rate. With reference to Lerner's observation of the dependence of interest rates not only on the quantity of capital involved but also on investment rates, Lange thought that, if saving was ceded to individual consumers, accumulation rates could be made to reflect consumers' preference. His 1936–8 model should be improved in this way, he said.

In his two public lectures delivered in Chicago in 1942 on 'The Economic Operation of a Socialist Society' (1975), Lange tacitly dropped what was perhaps the chief feature of his classical model, namely, the central authority's prerogative of setting and reviewing prices as a road towards equilibrium. He made only a passing remark about such a possibility, and only in reference to *future* prices the centre may impose on production managers in order to ensure stable forecasting (which is as a rule erratic in capitalist economy).

But perhaps the greatest change in his concept of the desired shape of socialism can be found in his above-mentioned article on economic foundations of democracy in Poland (1943b). The title alone shows that a commitment to furnish solid economic foundations for 'Poland's democratic order' was the point of departure in designing future political transformations. In that article, Lange envisaged the socialization only of key industries (which necessarily include banks and transport).This would put an end to the power of 'the socially irresponsible monopolistic capitalism'. Having said this, he cautions that care should be taken to prevent the socialized key industries from becoming a foundation for 'an equally dangerous' threat to democracy in the form of too much economic power being concentrated in the state bureaucracy along with privileges arising from this.

But private farms, crafts shops and minor but also medium-sized industries were all to remain areas of private initiative and enterprise. So broad a field of action for private entrepreneurship was, on the one hand, to be one foundation of democracy, and, on the other, it was to preserve 'the kind of flexibility, pliability and adaptiveness that private initiative alone can achieve'. This is the reason for which the development of private sector is to be one of the chief guidelines for the socialized financial policy. The private sector then appears to have been a permanent element of the new model Lange proposed for Poland.

This proposal had its counterpart for the United States in the lengthy essay written with Abba P. Lerner on a democratic programme for full employment (1944b).

The changes in Lange's views of socialist economy during the war years were evidently so substantial that they could be used to compose from them an alternative version of a market socialism, compared with which his classical model can indeed be described as 'quasi-centralistic' (Pryor 1985). The extent of those changes may have been the reason why he dropped his previous plan to revise his classical study:

> The essay is so far removed from what I would write on the subject today that I am afraid that any revision would produce a very poor compromise, unrepresentative of my thoughts. Thus, I am

> becoming inclined to let the essay go out of print and express my present views in entirely new form. I am writing a book on economic theory in which a chapter will be devoted to this subject. This may be better than trying to rehash old stuff. (Letter to M. Harding, 25 May 1945: 1986, p. 553.

Towards a Major Synthesis

Lange's lifelong ambition to produce a synthesis can be seen to have differed in scope, so that a 'minor' and a 'major' synthesis can be distinguished in it. His earliest endeavours included an attempt to incorporate the Marshall's method of partial equilibrium into the general equilibrium theory developed by the mathematical school (1932). In later years, he wrote a series of studies commenting on various aspects of the Keynesian theory to include it in and reduce to a particular case of general equilibrium theory.

Several times during his life Lange prepared himself to create his major synthesis. He did have the indispensable background for such a job, not only on account of his economic versatility (he was intimately familiar with all the main currents and schools in economic theory, and with the 'three economic worlds') but also because he felt at home in several other disciplines such as statistics and econometrics, history and sociology, praxeology and cybernetics.

The first outline for a major synthesis came in his article 'Marxian Economics and Modern Economic Theory' (1935). His chief argument was that these two currents are in fact complementary. Their advantages and drawbacks arose from the different specific tasks each of them was supposed to do. Marxian economics was designed to furnish the revolutionary movement with guidance for rational policies, defining as it did the lines and limitations of the evolution of capitalism. Modern economic theory, for its part, was expected to provide a foundation for capitalist management. But equilibrium theory, which was designed to serve precisely this purpose, was actually universal in character, so after some adaptation it could be used for day-to-day management of a socialist economy, a job Marxist economics was ill-suited to do. For some time Lange thought his synthesis should be based on marginalist economics, the categories of which seemed even useful for presenting problems of class structure. Clinging to 'Marxist semantics' was to him a sign of traditionalism and conservative attitudes.

In the late 1950s, he began to work on a three-volume treatise on political economy that would rest on two tiers – historical materialism and the principle of rationality. On a lower level of abstraction he attempted, rather unsuccessfully, to synthesize Marxian political economy with the neoclassical economics. He managed to finish the first volume (1959, 1963) on scope and method of economics and half of the second one (1966, 1971a). However, Poland was at that time only at the beginning of shedding its isolation straitjackets and thus of rapidly changing political, ideological and scientific perspectives and possibilities. That is why, only four years after the publication of the first volume, Lange came to a conclusion that, after having written the two next ones, it would need a substantial reworking.

From Idea to Reality

Having returned to Poland after the Second World War Lange gave an entirely new expression to his view of socialist economy. But by an ironic twist of history (to which he was fond of referring) he articulated his new approach only when his views changed in an entirely different direction from what he was pursuing during the wartime: namely, Lange embarked on the search for a rationale for the command- type economy and subsequently for ways of reforming it.

The evolution of Lange's views of socialism in the post-war years is much harder to follow because he became so deeply involved in political activity. Not only the form but also the substance of his views was often influenced by tactical considerations and by the changing scope of freedom of expression accorded to scholars in social science. The freedom was broad prior to 1948, virtually extinct in the early 1950s, considerable in the latter half of the 1950s, and gradually curtailed later on.

The main change in Lange's theoretical approach was that he switched over from a micro to a macroeconomic approach. Whereas he had previously based his argument on the general equilibrium theory, after 1945 he relied on a Marxian reproduction model. The new approach was first presented in the report he submitted to the International Statistical Conference (1947) on practical economic planning and optimal resource allocation. In this report he tried to confront eastern European economic practices with welfare economics. His point was that the centre's main decisions resulted from a desire to industrialize the country as rapidly as possible. The economic successes those countries had scored up to then were due to full employment and to the liquidation of private monopolies, which worked as powerful checks on their national economies in the past. Economic choices were a second-rate matter in the period of reconstruction, but as those countries were moving into a phase of development more sophisticated choices may have to be made. Marginal analysis may in such events prove useful, provided it is carried out in categories adequately reflecting reality. Although, Lange talked about practical planning in descriptive rather than theoretical terms and although he did not reject marginal analysis, F. Perroux said:

> Je note que le théoriciénsocialiste a complétement changé de méthode. Il a autrefois essayé de montrer qu'une économie socialiste peut fonctionner à peu près comme isolée des unités économie de marché, sur la base de calcul. . . . Il fonde aujourd'hui sa thèse sur les macro-décisions de l'Etat. Il le fait paradoxalement au moment précisément où tout le monde est d'accord surla necessité du 'breakdown of the aggregate quantities'. (1947, p. 172.

The new theoretical approach was given more clear-cut contours in a booklet (1953) in which Lange commented on Stalin's famous work on socialist economy in the USSR. The reasons for which Lange wrote that book, in which he extolled the Stalin work as 'a momentous event in the history of science with far-reaching practical consequences', are somewhat puzzling. He did it, probably, for two reasons. First, he was convinced that the Stalin work marked a turn from economic voluntarism towards respect for the inexorable laws governing economic life, towards a rehabilitation of efficiency and greater consideration of social needs. Indeed, the first studies written by Polish theorists who later became known as revisionists did find some support in Stalin's work.

The second reason that prompted him to write this booklet must have been his view of the evolution the Communist economies were undergoing due to industrialization. He believed that not only the Stalinist terror but also the main body of practical devices applied then, as well as the functioning of the economy itself at that time, were all determined by political considerations, specifically by militarization and the forceful industrialization bid (1943c). Lange often defined the centralistic command model as wartime economy. But he hoped that industrialization, with the subsequent emergence of an educated working class and socialist intelligentsia, creates a good social base for democracy and decentralization of management. Presuming that industrialization entailed democratization, he believed the future of the 'Polish economic model' depended on how mature and experienced society will be. This is why he was unwilling 'to design any new model from behind the desk'. In 1956–7 he refused to give his permission for the publication of an already finished translation of his classical work of 1936–8 because he did not want to lend his support to the 'socialist free-marketers'. But it is unclear whether he regarded the market-oriented model of socialism as premature or as invalidated by the progress made in economic theory and practice (1967).

Late in his life, cybernetics and mathematical programming became his fascination. Using the theory of systems self-regulation and self-control, Lange gave an interpretation of the chief categories, wholes and parts, of dialectical materialism ([1962] 1965a). He also wrote an introduction to economic cybernetics (1965b), and to the theory of optimal decisions (1971b). This fascination was born from a belief in a great role of the computer as a most powerful device for central planning (sometimes called 'computopia'). The strongest expression of this fascination contains his last publication on *The Computer and the Market* (1967). Recalling his polemics with Hayek and Mises, he confesses, that:

> Were I to rewrite my essay today my task would be much simpler. My answer to Hayek and Robbins would be: so what's the trouble? Let us put the simultaneous equations on an electronic computer and we shall obtain the solution in less than a second. The market process with its cumbersome *tatonnements* appears old-fashioned. Indeed, it may be considered as a computing device of the pre-electronic age.

It is rather obvious that such a view sharply contradicts his strong attachment to Marxian economics as economic sociology regarded by him as a seminal step in explaining a structure and evolution of capitalism.

This was, however, one side of his views. The other one, expressed rather in private communications, stems from his everyday observation and was truly pessimistic. Above we mentioned his opinion about the prime importance of incentives for the growth of productive forces termed by him as a true meaning of rationality. A couple of months before his death he did appreciate sociological factors of economic development: 'Poland became a completely parochial country. It is going to become the Portugal of the socialist block. The sociological setting generates an enduring stagnation, while an "explosive" solution of her problems stands no chance of success (nor does it seem really desirable). A change, if it comes, may be touched off by external developments, namely when Poland falls too far behind the capitalist world and the socialist world' (O. Lange's letter to T. Kowalik of 19 February 1965, in his possession).

Even if a comparison of Poland with the Salazar-time Portugal may be shocking, Lange's prophecy has proved to be quite realistic. Fifteen years later an 'explosive solution' in a form of a ten-million mass movement, 'Solidarity', brought first a lot of hopes, ended with martial law, but in another ten years Poland entered upon a track of peaceful dismantling of a Communist system. Theoretically, this could have opened a freedom for a democratic choice of socio-economic system, based if not on Lange's classical model literally then on his general democratic and egalitarian principles. It happened to the contrary. A wild form of capitalism emerged. Ronald Reagan, Margaret Thatcher, Milton Friedman and F.A. Hayek became prophets. The works of Oskar Lange and another eminent economist, Michał Kalecki, were rejected, their followers marginalized. At least five Polish politically engaged historians accused Lange of being a secret agent for the Soviet Union during his stay in the USA, although 13 volumes of declassified FBI documents clearly contradict this slander.

The Post-Langean Concepts of Market Socialism

Not all Western intellectuals have treated the sudden collapse of the Soviet bloc as a final victory of liberal capitalism. Even the *Washington Post* published (on 14 January 1990) an article entitled 'In Eastern Europe Social Democracy – not Capitalism of "1984" is winning'. Some of them saw the possibility of creating a new economic system, which would not simply emulate Western-type capitalism, and elaborated proposals using some of Lange's ideas as a starting point.

One of the first of them was Joseph Stiglitz, who as early as spring 1990 sent the following remarkable message to the post-Communist countries:

> The answer that socialism provided to the age-old question of the proper balance between the public and the private can now (. . .) be seen to have been wrong. But if it was based on wrong, or at least incomplete, economic theories (. . .) it was also based on ideals and values many of which are eternal. It represented a quest for a more humane and a more egalitarian society (. . .). As the former socialist countries embark on their journey, they see many paths diverging. There are not just two roads. Among these there are many that are less traveled by – where they end up no one yet knows. One of the large costs of the socialist experiment of the past seventy years is that it seemed *to foreclose exploring many of the other roads.* As the former socialist economies set off on this journey, let us hope that they keep in mind not only the narrower set of economic questions that I have raised (. . .) but the broader set of social ideals that motivated many of the founders of the socialist tradition. Perhaps some of them will *take the road less traveled by, and perhaps that will make all the difference, not only for them, but for the rest of us as well*. (Stiglitz 1990, p. 70, 1994, p. 279, emphasis added)

Stiglitz was very critical about the Lange–Lerner model of market socialism as based on wrong premises of the neoclassical paradigm. However, inspired by more general ideas of Lange, Michał Kalecki and the experience of Chinese gradual reforms, he suggested to the post-Communist countries several (for an American mainstream economist) very unconventional recommendations: not shock therapy as favoured by the IMF experts and particularly by Jeffrey Sachs, but evolutionary systemic changes; not market versus the state, but a search for the proper balance between market and government, the private and the state sector; not imitation of Anglo-Saxon capitalism, but a search for people's capitalism. He stressed that the post-Communist countries had most probably a chance to create a more egalitarian socio-economic system than any Western country. It was to be a mixed (market-cum-state) economy striving for social justice.

The efforts of a British philosopher and political scientist David Miller (1989) went in a different direction. Trying to create the theoretical foundations of market socialism, he explicitly says that his model would involve even 'more extensive use of markets' then the classical model of Oskar Lange.

Several economists also presented different concepts as alternatives to capitalism directly referring to some of Lange's ideas. The best known among them is John E. Roemer's (1994) proposal. Searching for an alternative system, which would be at least as efficient as present-day capitalism, he proposes to organize corporations in groups, operated according to the rules of the Japanese corporations called *keiretsu* with main banks crediting and monitoring them. Corporations would have to transfer after-tax profits to a state agency which would distribute it among all citizens as social dividend. This idea of a social dividend borrowed from Lange would be the main socialist feature of Roemer's model, which was nevertheless criticized as closer to capitalism than to socialist ideas.

Another American economist, James A. Yunker (1992), declared himself to be an enthusiast of Lange, not so much as an author of a classical model of socialism, but rather as a socialist thinker and particularly as a pioneer of reconciliation of conflicting theories. In this vein, he was arguing at the beginning of the 1990s for 'East–West ideological convergence' (Yunker 1993) based on his 'pragmatic market socialism' presented in many publications. He took over from his master only certain ideas, such as the social dividend, the interest rate as the main regulator of investment, and the scope of public ownership to be limited to firms where management was separated from ownership. But in other respects Yunker's model was quite far from its original inspiration. The institutional crux of his concept was to be – as he writes – the Bureau of Public Ownership, which would take over all rights inherent in stocks, bonds and other financial instruments owned by private households. The operation of this public sector would be based on institutional investing, which would proceed much as it does in present-day capitalism.

Contrary to Stiglitz, both Roemer's and Yunker's models are based on fully fledged market mechanism and the neoclassical paradigm.

Different character of a book is that by Włodzimierz Brus (Oxford) and Kazimierz Łaski (Vienna) (1989), both Polish emigrants as a result of the anti-Semitic campaign of March 1968. Earlier, while in Poland, Brus was a close collaborator of Lange and an eminent and very influential reform economist in the central European debates. Already in the beginning of the 1980s he became sceptical about the viability of market socialism. As a result of the analysis of its theoretical foundations and particularly a summary of the outcomes of reforms in the Soviet bloc, Yugoslavia and China, Brus and Łaski are inclined to abandon the very concept of socialism meant as an economic organization radically different from capitalism. They do not reject, however, socialist ideals, but see them as possibly realized rather in Scandinavian-type reforms of capitalism. After the collapse of Communism they saw some sort of market socialism rather as a necessary stage of transition to a new socio-economic system, when a coexistence of public and private ownership will be tolerated.

Needless to say, in all countries of central and eastern Europe the above-mentioned concepts,

even as cautious and moderate as that of Brus and Łaski, fell on deaf ears.

See Also

- ▶ Decentralization
- ▶ Economic Calculation in Socialist Countries
- ▶ Efficient Allocation
- ▶ Planning

Selected Works

1928a. Koniunktura w zyciugospodarczymPolski 1923–27 [Business performance in Poland's economic life]. In *Przewrotywalutowe i gospodarczepowielkiejwojnie* [Currency and economic upheavals after the Great War]. Kraków.

1928b. *Socjologia i ideespołeczne Edwarda Abramowskiego* [Sociology and social ideas of Edward Abramowski]. Kraków.

1929. 'Wrastanie w socjalizm' czy nowa faza kapitalizmu? ['Growing into' socialism, or a new phase of capitalism?]. *Robotniczy Przegląd Gospodarczy* 3: 69–74.

1931a. *Statystyczne badanie koniunktury gospodarczej* [Statistical investigation of the business cycle]. Kraków.

1931b. Kryzys socjalizmu [Crisis of socialism]. In *Historia, ideologia, zadania Związku Niezależnej Młodzieży Socjalistycznej* [History, ideology, goals of the Association of Independent Socialist Youth]. Kraków.

1932. Die allgemeine Interdependenz der Wirtschaftsprognosen und die Isolierungsmethode. *Zeitschrift für Nationalökonomie* 4(1).

1933. Odkryzysu do stabilizacji kapitalizmu [From crisis to the stabilization of capitalism]. *Płomienie* 17/18: 218–225.

1934. (With M. Breit.) Droga do socjalistycznej gospodarki planowej [The road to the socialist planned economy]. In *Gospodarka–polityka–taktyka–organizacjasocjalizmu* [Economy–polity–tactics–organization of socialism]. Warsaw.

1935. Marxian economics and modern economic theory. *Review of Economic Studies* 2(3): 189–201.

1936a. On the economic theory of socialism, Part I. *Review of Economic Studies* 4(1): 53–71.

1936b. The place of interest in the theory of production. *Review of Economic Studies* 3(3): 159–192.

1937a. On the economic theory of socialism, Part II. *Review of Economic Studies* 4(2): 123–142.

1937b. Professor Knight's note on interest theory. *Review of Economic Studies* 4(4): 231–235.

1938a. The rate of interest and the optimum propensity to consume. *Economica* 5: 12–32.

1938b. (With F.M. Taylor.) *On the economic theory of socialism.* With an introduction by Benjamin Lippincott. Minneapolis: University of Minnesota Press.

1941a. *Review of J. Schumpeter.* Business cycles: A theoretical, historical and statistical analysis of the capitalist process. *Review of Economic Statistics* 23(4): 190–193.

1941b. *Review of M. Kalecki.* Essays in the theory of economic fluctuations. *Journal of Political Economy* 49: 279–285.

1942a. The foundations of welfare economics. *Econometrica* 10: 215–228.

1942b. Say's law: A criticism and restatement. In *Studies in mathematical economics and econometrics,* ed. O. Lange et al. Chicago: University of Chicago Press.

1942c. Review of H.D. Dickinson. Economics of socialism. *Journal of Political Economy* 50(2): 299–303.

1943a. The theory of the multiplier. *Econometrica* 11: 227–245.

1943b. Gospodarcze podstawy demokracji w Polsce [Economic foundations of democracy in Poland]. In *Ku gospodarceplanowej* [Towards a centrally planned economy]. London.

1943c. *Working principles of the soviet economy.* New York: Russian Economic Institute.

1944a. *Price flexibility and employment.* Bloomington: Principia Press.

1944b. (With A.P. Lerner.) The American way of business. In *Problems in American life: Teaching aids for the social studies.* Washington, DC: National Council for the Social Studies.

1947. The practice of economic planning and the optimum allocation of resources. In *Proceedings of the international statistical conference,* vol. 5, Washington, DC. Published in *Econometrica* 17 (1949), 166–171.

1953. *Zagadnienia ekonomiipolitycznej w swietle-pracy J. Stalina 'Ekonomiczneproblemy-socjalizmu w ZSRR'* [Problems of political economy in the light of J. Stalin's work, 'Economic Problems of Socialism in the USSR']. Warsaw.

1957. *Dlaczego kapitalizm nie potrafi rozwiazać problem krajów gospodarczo zacofanych* [Why capitalism is unable to solve the problems of backward countries]. Warsaw.

1959. *Introduction to econometrics.* Oxford/London: Pergamon Press.

1963. *Political economy, vol. 1: General problems*. Oxford/London: Pergamon Press.

1965a. *Wholes and parts: A general theory of system behaviour*. Oxford/London: Pergamon Press.

1965b. *Introduction to economic cybernetics.* Warsaw: Państwowe Wydawnictwo Naukowe.

1966. *Ekonomiapolityczna,* vol. 2. Warsaw: Państwowe Wydawnictwo Naukowe.

1967. The computer and the market. In *Socialism, capitalism and economic growth: Essays presented to M. Dobb,* ed. C.H. Feinstein. Cambridge: Cambridge University Press.

1970. *Papers in economics and sociology.* Trans. P.F. Knightsfield. Oxford: Pergamon Press.

1971a. *Political economy,* vol. 2. Warsaw: Państwowe Wydawnictwo Naukowe.

1971b. *Optimal decision: Principles of programming*. Oxford: Pergamon Press.

1973. *Dziela* [Works], vol. 1. Warsaw: Państwowe Wydawnictwo Ekonomiczne.

1975. *Dziela* [Works], vol. 3. Warsaw: Państwowe Wydawnictwo Ekonomiczne.

1986. *Dzieła* [Works], vol. 8. Warsaw: Państwowe Wydawnictwo Ekonomiczne.

1994. *Economic theory and market socialism: Selected essays of Oskar Lange*, ed. T. Kowalik. Aldershot: Edward Elgar Publishing Limited.

2003. (With M. Breit.) The way to the socialist planned economy (1934). Trans. J. Toporowski. *History of Economics Review* 37: 51–70.

Bibliography

Bergson, A. 1967. Market socialism revisited. *Journal of Political Economy* 75: 655–673.

Brus, W., and K. Łaski. 1989. *From Marx to the market: Socialism in search of an economic system*. Oxford: Clarendon Press.

Burk, A. [pen-name of Abraham Bergson]. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52: 310–334.

Chilosi, A. 1986. Self-managed market socialism with 'free mobility of labor'. *Journal of Comparative Economics* 10: 237–254.

Chilosi, A. 2005. The Lange–Breit model, the right to employment and the outside option. In *Oskar Lange a współczesno*. Warsaw: Polskie Towarzystwo Ekonomiczne.

Graaff, J. de V. 1957. *Theoretical welfare economics.* Foreword by P.A. Samuelson. Cambridge: Cambridge University Press.

Keynes, J.M. 1973. *Collected writings, vol. XIV: The general theory and after.* London: Macmillan.

Kowalik, T. 1970. Oskara Langego Wczesne modele socjalizmu [O. Lange's early models of socialism]. *Ekonomista* 5: 965–1000.

Kowalik, T. 1974. Zurklassischem Modell des Sozialismus. In *Sozialismus, Geschichte und Wirtschaft: Festschrift fur Eduard Marz*. Vienna: Europaverlag.

Kowalik, T. 1984. Review of A. Nove. The economics of feasible socialism. *Contributions to Political Economy* 3: 91–97.

Lavoie, D. 1985. *Rivalry and central planning: The socialist calculation debate reconsidered*. Cambridge: Cambridge University Press.

Lerner, A.P. 1936. A note on socialist economics. *Review of Economic Studies* 4: 72–76.

Miller, D. 1989. *Market, state and community. Theoretical foundations of market socialism.*Oxford: Clarendon Paperbacks.

Perroux, F. 1947. Comments on Lange's paper. In *Proceedings of the international statistical conference*, vol. 5. Washington, DC. Published in *Econometrica* 17 (1949), 172–178.

Pryor, F.L. 1985. *A guidebook to the comparative study of economic systems*. Englewood Cliffs: Prentice-Hall.

Roemer, J.A. 1994. *A future for socialism*. Cambridge, MA: Harvard University Press.

Stiglitz, J. 1990. *Whither socialism? Perspectives from the economics of information*. Mimeo: Stockholm.
Stiglitz, J. 1994. *Whither socialism? Perspectives from the economics of information*. Cambridge, MA: MIT Press.
Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19: 1–8.
Toporowski, J. 2003. MarekBreit and Oskar Lange's financial model of a socialist economy. *History of Economics Review* 37: 41–50.
Yunker, J.A. 1992. *Socialism revised and modernized: The case for pragmatic market socialism*. New York: Praeger Publishers.
Yunker, J.A. 1993. New prospects for East–West ideological convergence: A market socialist viewpoint. *Coexistence* 30: 237–267.

Lange–Lerner Mechanism

Tadeusz Kowalik

This is a designation commonly used to denote a market-oriented socialism model devised by Lange, who later amended it after public discussion with Lerner. The first, fundamental part of Lange's study was published together with A.P. Lerner's (1936) critical remarks in the same issue of the *Review of Economic Studies*, while the second part appeared together with Lange's reply to Lerner (1937). Later on, Lange made the changes necessary to publish his study (together with F.M. Taylor's essay) in book form (1938). The term is occasionally used in a less restricted sense, to bring out the similarity of Lange's and Lerner's views on other matters concerning market socialism.

The mechanism of socialist economy in the Lange–Lerner blueprint was based on the following assumptions. It has its institutional framework in the public ownership of means of production (for simplicity, the private sector is omitted) and in the free choice of consumption and employment (job and workplace), while consumer preferences – 'through demand prices' – are the all-decisive criterion of both production and resource allocation. Under these assumptions, an authentic market (in the institutional sense) exists for consumer goods and labour services. But prices of capital goods and 'all other productive resources except labour' are set by a Central Planning Board (CPB) as indicators of existing alternatives established for the purpose of economic calculation. So, apart from market prices, there are also 'accounting prices'. Both categories of prices are used by enterprise and industry managers, who are public officials, in order to make their choices.

Production managers in charge of individual enterprises or entire industries make autonomous decisions about what and how much should be produced and how it should be done, while prices are set as parameters outside the enterprises or industries. But since profit maximization has by definition ceased to be a direct goal of economic activity, to ensure that they can achieve effects close to those achieved in free-market economy, production managers must obey two rules. First, they must pick a combination of production factors under which average cost is minimized, and second, they must determine a given industry's total output at a level at which marginal cost is equal to product price. The first rule was expected to eliminate all less efficient alternatives. In combination with the second rule, insofar as it concerns plant managers, it performs the same function as the free-market economy desire to maximize profit. This leads to minimization of production costs. The second rule compels production managers to increase or cut the output of a whole industry in accordance with consumer preferences, which is a substitute for free entry in a free competitive economy.

These rules lead to an economic equilibrium by the trial-and-error method first described by Fred M. Taylor (1929). The CPB acts like an auctioneer, initially watching the behaviour of economic actors in reaction to a price system it picks at random or – perhaps the best solution – to the historically inherited prices. The behaviour of the system is measured by the movement of inventories of goods. If there is too much of some product at a given price, then its inventory grows, and vice versa. This is regarded as information that the product price should be cut or increased, respectively. This procedure is applied as many times as is necessary to reach equilibrium, providing that this process does in fact converge to the system of

equilibrium prices. Accounting prices, then, are objective in character, just like market prices in a competitive system, the difference being that in this case the CPB performs the role of the market. The same trial-and-error way toward equilibrium could also be applied in two other models of socialist economy, one providing for a decreased consumer influence on production programme, the other presupposing none at all.

In its extreme version, which for sociopolitical reasons Lange deems untenable, the model might provide no freedom of choice for either consumption or employment. Production plans would be decided by the CPB officials' scale of preferences. In such a version all prices are basically accounting prices. Consumer goods are rationed, while the place and kind of employment are imposed by command. If production managers keep to the above-mentioned rules, and if the CPB keeps to the parametric price system, then economic calculus is possible even in this version, while prices are not arbitrary but reflect the relative scarcity of factors of production.

There is an intermediate model, which provides for freedom of consumption decisions but only within a production plan established on the ground of CPB preferences. In this case, accounting prices of producer and consumer goods reflect the CPB's preference scale, while production managers would rely on them in their decision-making. Market prices for consumer goods would be set by supply and demand. But Lange rejects even this system as undemocratic, saying that the dual system of prices could be applied only when there is widespread agreement that checking the consumption of some products (say, alcohol) while promoting the consumption of other goods (say, cultural services) is in the public interest.

But the CPB might conceal its preferences and resort to rationing production goods and resources. Society can defend itself against such practices by creating a supreme economic court which would be entitled to declare any unconstitutional CPB decision as null and void. In Lange's view, any decision introducing rationing would be unconstitutional.

Interestingly, Lange rejects these two versions of socialist economy on account of the potential hazards they carry for democracy, and says not a word about democracy's possible link with economic efficiency.

Lange considers the distribution of national income in three aspects.

Wages would be differentiated by seeking a distribution of labour services that would maximize society's wealth in general. This happens when differences in marginal disutility of work in different trades and workplaces are offset by wage differences. Wage differentials can be treated as converses of prices paid by employees for differing work conditions, as a simplified form of buying free time, safety or pleasant work (which is easy to imagine assuming that all employees get the same earnings but pay different prices for doing different jobs; the easier and safer a given job, the more one has to pay for it). In this sense, the wage differentiation rule can be brought into harmony with egalitarianism.

Apart from wages paid by employees, each consumer is paid a public dividend as his or her share of capital and natural resources. At first Lange was inclined to distribute such dividends proportionally to wages. But as Lerner pointed out that such a policy would impart added attractiveness to the hardest jobs, Lange changed his mind, saying there should be no link between procedures for public dividend distribution and wage differentials.

The distribution of national income between consumption and accumulation, said Lange, would not be arbitrary when only consumers' individual savings decide the rate of accumulation. But if savings are 'corporately' determined – and Lange at first thought that was typical of a socialist economy – then there would be no way of preventing the CPB from being at least partly arbitrary in its decisions.

Emphasizing that resource allocation is guided by formally analogous rules in both socialist and free competitive economies, Lange argued that real allocation in socialism would be different from and more rational than that in capitalism. In his static analysis, he considered the following factors as decisive in judging the relative performance of the two systems. Greater equality of income distribution enhances society's well-being (in the subjective sense, that is, as a sum

total of individual satisfactions). Second, socialist economy makes allowances in its calculus for all the services rendered by producers and for all the costs involved, while a private entrepreneur does not care for benefits that do not flow into his own pocket nor for costs he does not have to pay: 'Most important alternatives, like life, security, and health of the workers, are sacrificed without being accounted for as a cost of production' (1938, p. 104).

Even the possible flaws that Lange conceded might appear in a socialist economy, such as the arbitrary setting of the rate of accumulation or the danger of bureaucratization of economic life, would be milder than under capitalism, he argued.

But the ultimately decisive economic argument in favour of socialism, Lange believed, was the general waste and endogenous tendency toward stagnation generated by modern capitalism's monopolistic tendencies. This question, though, goes beyond the scope of the often-criticized static analysis underlying Lange's classical model. Leaving aside the now enormous critical literature, let us try to answer the question of what Lange himself saw as his model's limitations.

Lange anticipated possible charges by critics in the second part of his study, in his discussion of 'The Economist's Case for Socialism':

> The really important point in discussing the economic merits of socialism is not that of comparing the equilibrium position of a socialist and of a capitalist economy with respect to social welfare. Interesting as such a comparison is for the economic theorist, it is not the real issue in the discussion of socialism. The real issue is *whether the further maintenance of the capitalist system is compatible with economic progress* (1938, p. 110).

But as he develops this general idea, Lange clearly uses an asymmetrical kind of argument. Having presented free competitive capitalism as the system which generated 'the greatest economic progress in human history', Lange proceeds to show (among other things, by referring to Keynes) that the source of that progress is drying up because of the progressive concentration and monopolization of production. His main point is that corporations, which are capable of controlling the market, attempt to avoid losses due to capital depreciation caused by innovation, and hence they try to check progress in technology. Neither a return to free competition nor government control can effectively eliminate this tendency. The only effective solution, then, is the socialization of big capital, the introduction of socialism.

But will socialism ensure rapid technical progress? Will the abolition, via socialization, of capitalist monopolies' well-known tendency to check technological progress automatically dismantle all the barriers to innovation? Or will it amount to substituting new barriers for old? Will the two rules for managers be sufficient to guarantee the adoption of state-of-the-art production techniques? In his classic study, Lange never even asked such questions and only much later did he become aware of them.

Toward the end of his life (in a letter to the present writer dated 14 August 1964), Lange wrote:

> What is called optimal allocation is a second-rate matter, what is really of prime importance is that of incentives for the growth or productive forces (accumulation and progress in technology). This is the true meaning of, so to say, 'rationality'.

It seems that he must have lacked the indispensable tools to solve this question or even to present it in detail.

See Also

- ▶ Control and Coordination of Economic Activity
- ▶ Decentralization
- ▶ Economic Calculation in Socialist Countries
- ▶ Efficient Allocation
- ▶ Market Socialism
- ▶ Planning
- ▶ Socialist Economies

Bibliography

Lange, O. 1936–7. On the economic theory of socialism. Pts I–II. *Review of Economic Studies* 4, Pt I, October 1936, 53–71; Pt II, February 1937, 123–142.

Lange, O., and F.M. Taylor. 1938. *On the economic theory of socialism*. Ed. and with an introduction by Benjamin E. Lippincott. Minneapolis: University of Minnesota Press. Reprinted, New York: McGraw-Hill, 1964.

Lerner, A.P. 1936. A note on socialist economics. *Review of Economic Studies* 4: 72–76.
Taylor, F.M. 1929. The guidance of production in a socialist state. *American Economic Review* 19: 1–8.

Lardner, Dionysius (1793–1859)

Robert B. Ekelund Jr.

Keywords
Jevons, W. S.; Lardner, D.; Mathematical economics; Price discrimination; Profit maximization; Spatial economics

JEL Classifications
B31

Scientific popularizer and railway economist, Lardner was born in Dublin on 3 April 1793 and died on 29 April 1859. He was educated at Trinity College, Dublin, between 1817 and 1827 and is probably best known for his *Cabinet Cyclopaedia* of 133 volumes, published between 1829 and 1849. Although Lardner's series was graced by a number of distinguished contributors, he was satirized in the scientific community as 'Dionysius Diddler'. An astronomer as well as an essayist on numerous scientific topics, Lardner often took side trips into other fields. He studied railway engineering in Paris, and was probably well acquainted with the econo- engineering work at the Ecole des Ponts et Chaussées at a time when Jules Dupuit was actively pursuing economic topics. His sole work relating to economics, *Railway Economy* (1850), was filled with the kind of factual work and analysis being undertaken by the French engineers and by an American pupil of the Ecole, Charles Ellet. Lardner's work caught the eye of W.S. Jevons, who claimed that a reading of *Railway Economy* in 1857 led him to investigate economics in mathematical terms.

There is little doubt that Lardner's book contains important and creative insights into economic theory. An authority on Belgian railroads of the time, Lardner drew up a vast array of facts to develop a theory of the railway firm's costs and revenues. His theory of profit maximization derived from 'empirical' firm's costs and revenues may be set out graphically (see Fig. 1).

The railway tariff, which Lardner identified as the independent variable, is displayed on the horizontal axis of the figure while total cost and receipts are measured on the horizontal. The total cost curve shows costs increasing as the tariff is lowered. At a prohibitive tariff Ox, that is, where no traffic would be transported, costs are some positive amount. Fixed costs, which exist whether traffic is carried or not, are an amount xL. As the tariff is lowered, increases in traffic carried cause total costs to increase until they reach maximum at a zero tariff. Both fixed and variable components of cost, then, are considered by Lardner.

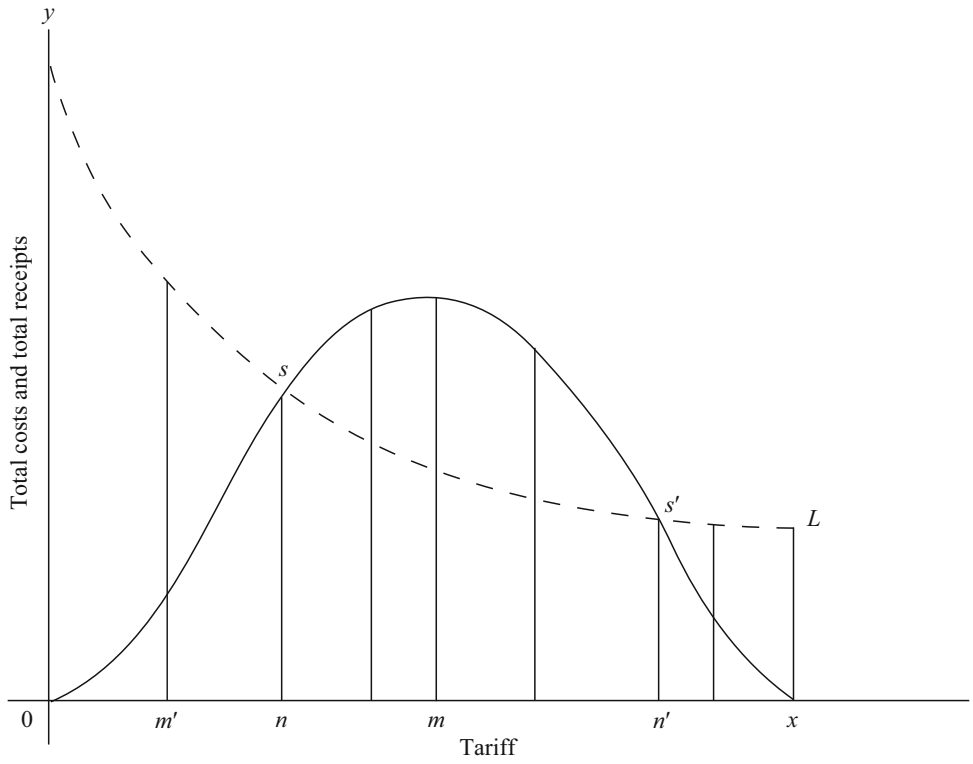
Lardner formalized his conception of total receipts in the following terms. If, with Lardner, we let

r = the tariff imposed per mile on each ton of goods carried;
D = the average distance in miles to which each ton of goods is carried;
N = the number of tons booked, and;
R = the gross receipts from goods transport, then total receipts may be expressed as

$$R = NDr.$$

As the tariff is lowered from Ox, the average distance of each ton carried, D, and the number of tons booked, N, increase. With reference to Fig. 1, lowering the tariff from Ox causes receipts, R in Lardner's equation, to increase to some maximum mp. Tariff reductions below Om, however, cause total receipts to fall, so that at a tariff of zero, total receipts are zero (demand is inelastic for tariffs below Om).

Tariffs On' and On are 'break even' tariffs in Fig. 1 and, significantly, Lardner argued that the profit-maximizing tariff would fall somewhere between the break-even tariff On' and the revenue-maximizing tariff Om. In modern terminology Lardner identified, if implicitly, the profit maximizing *quantity* as being where marginal



Lardner, Dionysius (1793–1859), Fig. 1

cost equals marginal revenue. It is noteworthy that Lardner's analysis of profit maximization, which so impressed Jevons, is nowhere to be found in Jevons's writings.

In addition to a fine model of the profit-maximizing firm, Lardner presented a fairly complete theory of price discrimination related to location in his *Railway Economy*. Specifically, Lardner called for a reduction in long-haul rail rates and for the increase in short-haul rates in order to increase the aggregate profits of the railroad. The differing elasticities of demand for transport which made this discriminatory pricing structure possible were explained on the basis of spatially distributed demanders.

Selected Works

1850. *Railway economy*. Reprinted, New York: A.M. Kelley, 1968.

Bibliography

Ekelund, R.B. Jr., E.G. Furubotn, and W.P. Gramm. 1972. *The evolution of modern demand theory*. Lexington: Heath.

Hooks, D.L. 1971. Monopoly price discrimination in 1850: Dionysius Lardner. *History of Political Economy* 3: 208–223.

Robertson, R.M. 1951. Jevons and his precursors. *Econometrica* 19: 229–242.

Large Economies

John Roberts

Keywords

Contract curve; Convergence; Convexity; Cournot, A. A.; Large economies; Nonstandard analysis; Oligopoly; Richter's theorem

JEL Classifications
O1

Economists have often claimed that our theories were never intended to describe individual behaviour in all its idiosyncrasies. Instead, in this view, economic theory is supposed to explain only general patterns across large populations. The prime example is the theory of competitive markets, which is designed to deal with situations in which the influence of any individual agent on price formation is 'negligible'.

As in so many aspects of economics, Cournot (1838) was the first to make the role of large numbers explicit in his analysis. Cournot provided a theory of price and output which, as the number of competing suppliers increases without bound, asymptotically yields the competitive solution of price equals marginal and average cost. However, for any given finite number of competitors, an imperfectly competitive outcome results.

It took over a century for Cournot's insights on the role of large numbers to be fully appreciated. Edgeworth (1881) argued the convergence of his contract curve as the economy grew, and increasing numbers of authors assumed that the number of agents was 'sufficiently large' that each one's influence on quantity choices was negligible, but it was not until the contributions of Shubik (1959) and Debreu and Scarf (1963) to the study of the asymptotic properties of the core that the number of agents took a central role in economic analysis.

The crucial step in this line of analysis was taken by Aumann (1964). Arguing that, in terms of standard models of behaviour, an individual agent's actions could be considered to be negligible only if the individual were himself arbitrarily small relative to the collectivity, Aumann modelled the set of agents as being (indexed by) an atomless measure space. In this context, an individual agent corresponds to a set of measure zero, while aggregate quantities are represented as integrals (average, per capita amounts). Then changing the actions of a single individual (or any finite number) actually has no influence on aggregates.

The non-atomic measure space formulation brings three mathematical properties that have proven important. The first is that it provides a consistent modelling of the notion of individual negligibility: only in such a context is an individual truly able to exert no influence on prices. Thus, this model correctly represents the primary reason for appealing to 'large numbers': in it, competitive price-taking behaviour is rational.

Moreover, this individual negligibility, when combined with an assumption that individual characteristics are sufficiently 'diffuse', means that discontinuities in individual demand disappear under aggregation (Sondermann 1975).

The second property is that a (non-negligible) subset of agents drawn from an economy with a non-atomic continuum of agents is essentially sure to be a representative sample of the whole population. This property has proven crucial in the literature relating the core and competitive equilibrium. (See Hildenbrand 1974, for a broad-ranging treatment of these issues.) It is also used in showing equivalence of core and value allocations (Aumann 1975).

The other important property of the non-atomic continuum model is the convexifying effect. Even though individual entities (demand correspondences, upper-contour sets, production sets) may not be convex, Richter's theorem implies that the aggregates of these are convex sets when the set of agents is a non-atomic continuum. This property yields existence of competitive equilibrium in large economies even when the individual entities are ill behaved and no 'diffuseness' is assumed.

In the non-atomic continuum modelling, the individual agent formally disappears. Instead, one has coalitions (measurable sets of agents), and an individual is formally indistinguishable from any set of measure zero. The irrelevance of individuals is made very clear in the model of Vind (1964), where only coalitions are defined and individual agents play no part. Debreu (1967) showed the equivalence of Vind's and Aumann's approaches. A further extension of this line is to consider economies in terms only of the distributions of individual characteristics and allocations in terms of distributions of commodities. The strengths of this approach are shown in Hildenbrand (1974).

This disappearance of the individual is intuitively bothersome: economists are used to thinking about individual agents being negligible, but not about individuals having no existence whatsoever. Brown and Robinson (1972) provided an escape from this dilemma by their modelling of a large set of agents via non-standard analysis. This approach gives formal meaning to such notions as an infinitesimal that had been swept out of mathematics and replaced by 'epsilon-delta' arguments. In interpreting non-standard models, one distinguishes between how things appear from 'inside the model' and what they look like from 'outside'. From outside, these models may have an infinity of (individually negligible, infinitesimal) agents, yet from inside each agent is a well-defined, identifiable entity. Using this mathematical modelling eases the interpretation of large economies and also allows formalization of some very intuitive arguments that otherwise could not be made. Unfortunately, the difficulties of mastering the mathematics of non-standard analysis have limited the number of economists using this approach.

While these formal models capture the essential intuition about the nature of economic behaviour of large economies, results obtained in this context should be of interest only to the extent that these models provide a good approximation to large but finite economies. This point was first emphasized by Kannai (1970), and its elaboration was the central issue confronting mathematical general equilibrium theory through the 1960s and early 1970s. The issue is one of continuity: in what sense are infinite economy models the limits of finite economies as the economy grows, and do the various constructs of interest (competitive or Lindahl allocations, cores, value allocations, and so on) of the finite economies approach those of the limit, infinite economies? These questions are extremely subtle. A good introduction to them is Hildenbrand (1974).

The study of the limiting, asymptotic properties of various economic concepts represents an alternative, more direct (but often less tractable) approach to large economy questions than does working with infinite economies. This line begins with Cournot's (1838) treatment of the convergence of oligopoly to perfect competition, the general equilibrium development of which has been a major focus of recent activity (see Mas-Colell 1982 and the references there). The work growing out of Edgeworth (1881) and Debreu and Scarf (1963) on the core-competitive equilibrium equivalence noted above also follows this line.

Once such convergence is established, the crucial question becomes that of the rate of convergence because asymptotic results are of limited interest if convergence is too slow. This question was first addressed for the core by Debreu (1975), who showed convergence at a rate of at least 1 over the number of agents.

A more direct approach to this issue of how large a market must be for its outcomes to be approximately competitive is to employ a model in which price formation is explicitly modelled. (Note that this is not a property of the Cournot or Arrow–Debreu analyses.) In a partial equilibrium context the Bertrand (1883) model of price-setting homogeneous oligopoly indicates that 'two is large', in that duopoly can yield price equal to marginal cost. Recent striking results in the same line for the double auction are due to Gresik and Satterthwaite (1985), who show that, even with individual reservation prices being private information, equilibrium under this institution can yield essentially competitive, welfare-maximizing volumes of trade with as few as six sellers and buyers.

This work is very heartening, for it tends to justify the profession's traditional reliance on competitive models which make formal sense only with an infinite set of agents. Another basis for optimism on this count comes from experimental work which shows strong tendencies for essentially competitive outcomes to be attained with quite small numbers. The further study of such institutions is clearly indicated.

See Also

- ▶ Non-standard Analysis
- ▶ Perfect Competition
- ▶ Shapley–Folkman Theorem

References

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Aumann, R.J. 1975. Values of markets with a continuum of traders. *Econometrica* 43: 611–646.

Bertrand, J. 1883. Theorie mathématique de la richesse sociale. *Journal des Savants* 48: 499–508.

Brown, D.J., and A. Robinson. 1972. A limit theorem on the cores of large standard exchange economies. *Proceedings of the National Academy of Sciences of the United States of America* 69: 1258–1260.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Debreu, G. 1967. Preference functions on measure spaces of economic agents. *Econometrica* 35: 111–122.

Debreu, G. 1975. The rate of convergence of the core of an economy. *Journal of Mathematical Economics* 2: 1–8.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Gresik, T., and M. Satterthwaite. 1985. *The rate at which a simple market becomes efficient as the number of traders increases: An asymptotic result for optimal trading mechanisms*. Discussion paper no. 641, Center for Mathematical Studies in Economics and Management Science, Northwestern University.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Kannai, Y. 1970. Continuity properties of the core of a market. *Econometrica* 38: 791–815.

Mas-Colell, A., ed. 1982. *Non-cooperative approaches to the theory of perfect competition*. New York: Academic Press.

Shubik, M. 1959. *Strategy and market structure: Competition, oligopoly, and the theory of games*. New York: Wiley.

Sondermann, D. 1975. Smoothing demand by aggregation. *Journal of Mathematical Economics* 2: 201–224.

Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5: 165–177.

Large Games (Structural Robustness)

Ehud Kalai

Abstract

In strategic games with many semi-anonymous players all the equilibria are structurally robust. The equilibria survive under structural alterations of the rules of the game and its information structure, even when the game is embedded in bigger games. Structural robustness implies *ex post* Nash conditions and a stronger condition of information-proofness. It also implies fast learning, self-purification and strong rational expectations in market games. Structurally robust equilibria may be used to model games with highly unspecified structures, such as games played on the web.

Keywords

Herding; Large games; Mixed-strategy equilibrium; Pure-strategy equilibrium; Purification; Rational expectations equilibrium; Structural robustness

JEL Classifications

C7

Earlier literature on large (many players) cooperative games is surveyed in Aumann and Shapley (1974). For large strategic games, see Schmeidler (1973) and the follow-up literature on the purification of Nash equilibria. There is also substantial literature on large games with special structures, for example large auctions as reported in Rustichini et al. (1994).

Unlike the above, this survey concentrates on the structural robustness of (general) Bayesian games with many semi-anonymous players, as developed in Kalai (2004, 2005). (For additional notions of robustness in game theory, see Bergemann and Morris 2005.)

Main Message and Examples

In simultaneous-move Bayesian games with many semi-anonymous players, all Nash equilibria are structurally robust. The equilibria survive under structural alterations that relax the simultaneous-play assumptions, and permit information transmission, revisions of choices, communication, commitments, delegation, and more.

Large economic and political systems and distributive systems such as the Web are examples of environments that give rise to such games.

Immunity to alterations means that Nash equilibrium predictions are valid even in games whose structure is largely unknown to modellers or to players.

The next example illustrates immunity of equilibrium to revisions, or being *ex post* Nash, see Cremer and McLean (1985), Green and Laffont (1987) and Wilson (1987) for early examples.

Example 1 *Ex post stability illustrated in match pennies* Simultaneously, each of k males and k females chooses one of two options, H or T. The payoff of every male is the proportion of females his choice matches and the payoff of every female is the proportion of males her choice *mis*matches. (When $k = 1$ this is the familiar match-pennies game.) Consider the mixed-strategy equilibrium where every player chooses H or T with equal probabilities.

Structural robustness implies that the equilibrium must be *ex post* Nash: it should survive in alterations that allow players to revise their choices after observing their opponents' choices. Clearly this is not the case when k is small. But as k becomes large, the equilibrium becomes arbitrarily close to being *ex post* Nash. More precisely, the Prob[some player can improve his payoff by more than ε *ex post*] decreases to zero at an exponential rate as k becomes large.

Example 2 *Invariance to sequential play illustrated in a computer choice game* Simultaneously, each of n players chooses one of two computers, I or M. But before choosing, with 0.50–0.50 i.i.d. probabilities, every player is privately informed that she is an I-type or an M-type. The payoff of every player is 0.1 if she chooses the computer of her type (zero otherwise) plus 0.9 times the proportion of opponents whose choices she matches. (Identical payoffs and prior probabilities are assumed only to ease the presentation. The robustness property holds without these assumptions.) Consider the favourite-computer equilibrium (FC) where every player chooses the computer of her type.

Structural robustness implies that the equilibrium must *be invariant to sequential play*: it should survive in alterations in which the (publicly observed) computer choices are made sequentially. Clearly this is not the case for small n, where any equilibrium must involve herding. But as n becomes large, the structural robustness theorem below implies that FC becomes an equilibrium in all sequential alterations. More precisely, the Prob[some player, by deviating to her non favorite computer, can achieve an ε-improvement at her turn] decreases to zero at an exponential rate.

The general definition of structural robustness, presented next, accommodates the above examples and much more.

Structural Robustness

A mixed-strategy (Nash) equilibrium $\sigma = (\sigma_1, \ldots, \sigma_n)$ of a one-simultaneous-move n-person strategic game G is structurally robust if it *remains an equilibrium* in *every* structural *alteration* of G. Such an alteration is described by an extensive game, $\mathcal{A}$, and for σ to remain an equilibrium in $\mathcal{A}$ means that *every adaptation* of σ to $\mathcal{A}$, $\sigma^{\mathcal{A}}$, must be an equilibrium in $\mathcal{A}$.

Consider any n-person one-simultaneous-move Bayesian game G, like the Computer Choice game above.

Definition 1 *A (structural) alteration of G* is any finite extensive game $\mathcal{A}$ with the following properties:

1. $\mathcal{A}$ *includes the (original) G-players*: The players of $\mathcal{A}$ constitute a superset of the G-players (the players of G).
2. *Unaltered type structure*: At the first stage of $\mathcal{A}$, the G-players are assigned a profile of types by the same prior probability distribution as in G. Every player is informed of his own type.
3. *Playing* $\mathcal{A}$ *means playing G*: with every final node of $\mathcal{A}$, z, there is an associated unique profile of G pure-strategies, $a(z) = (a_1(z), \ldots, a_n(z))$.
4. *Unaltered payoffs*: the payoffs of the G-players at every final node z are the same as their payoffs in G (at the profile of realized types and final pure-strategies $a(z)$).

5. *Preservation of original strategies*: every pure-strategy a_i of a G-player i has at least one $\mathcal{A}$ adaptation. That is, an $\mathcal{A}$ -strategy $a_i^{\mathcal{A}}$ that guarantees (w.p. 1) ending at a final node z with $a_i(z) = a_i$ (no matter what strategies are used by the opponents).

In the computer choice example, every play of an alteration $\mathcal{A}$ must produce a profile of computer allocations for the G-players. Their preferences in $\mathcal{A}$ are determined by their preferences over profiles of computer allocations in G. Moreover, every G-player i has at least one $\mathcal{A}$-strategy $I_i^{\mathcal{A}}$ (which guarantees ending at a final node where she is allocated I), and at least one $\mathcal{A}$-strategy $M_i^{\mathcal{A}}$ (which guarantees ending at a final node where she is allocated M).

Definition 2 *An $\mathcal{A}$ (mixed) strategy-profile, $\sigma^{\mathcal{A}}$, is an adaptation of* a G (mixed) strategy-profile σ, if for every G-player i, every $\sigma_i^{\mathcal{A}}$ is an $\mathcal{A}$-adaptation of σ_i. That is, for every G pure-strategy a_i, $\sigma_i(a_i) = \sigma_i^{\mathcal{A}}\left(a_i^{\mathcal{A}}\right)$ for some $\mathcal{A}$-adaptation $a_i^{\mathcal{A}}$ of a_i.

In the computer choice example, for a G-strategy where player i randomizes 0.20 to 0.80 between I and M, an $\mathcal{A}$ adaptation must randomize 0.20–0.80 between a strategy of the type $I_i^{\mathcal{A}}$ and a strategy of the type $M_i^{\mathcal{A}}$.

Definition 3 An equilibrium σ of G *is structurally robust* if in every alteration of G, $\mathcal{A}$, and in every adaptation of σ, $\sigma^{\mathcal{A}}$, the strategy of every G-player i, $\sigma_i^{\mathcal{A}}$, is best response to $\sigma_{-i}^{\mathcal{A}}$.

Remark 1 The structural robustness theorem, discussed later, presents an asymptotic result: the equilibria are structurally robust up to two positive numbers (ε,ρ), which can be made arbitrarily small as n becomes large. The notion of approximate robustness is the following.

An equilibrium is (ε, ρ)-structurally robust if in every alteration and every adaptation as above, Prob[visiting an information set where a G-player can improve his payoff by more than ε] $\leq \rho$. (ε-improvement is computed conditional on being at the information set. To gain such improvement the player may coordinate his deviation: he may make changes at the information set under consideration together with changes at forthcoming ones.)

For the sake of brevity, the next section discusses full structural robustness. But all the observations presented there also hold for the properly defined approximate counterparts. For example, the fact that structural robustness implies *ex post* Nash also implies that approximate structural robustness implies approximate *ex post* Nash. The implications of approximate (as opposed to full) structural robustness are important, due to the asymptotic nature of the structural robustness theorem.

Implications of Structural Robustness

Structural robustness of an equilibrium σ in a game G is a strong property, because the set of G-alterations that σ must survive is rich. The simple examples below are meant to suggest the richness of its implications, with the first two examples showing how it implies the notions already discussed (see Dubey and Kaneko 1984 for related issues).

Remark 2 ***Ex post Nash and being information-proof*** *G with revisions*, $\mathcal{GR}$, is the following n-person extensive game. The n players are assigned types as in G (using the prior type distribution of G and informing every player of his own type). In a first round of simultaneous play, every player chooses one of his G pure strategies; the types realized and pure strategies chosen are all made public knowledge. Then, in a second round of simultaneous play, the players again choose pure strategies of G (to revise their first round choices). The payoffs are as in G, computed at the profile of realized types with the profile of pure strategies chosen in the *second* round.

Clearly $\mathcal{GR}$ satisfies the definition of an alteration (with no additional players), and every equilibrium σ of G has the following $\mathcal{GR}$ adaptation, σ^{NoRev}: in the first round the players choose their pure strategies according to σ, just as they do in G; in the second round nobody revises his first round choice.

Structural robustness of σ implies that σ^{NoRev} must be an equilibrium of $\mathcal{GR}$, that is, σ is *ex post* Nash.

Moreover, the above reasoning continues to hold even if the information revealed between the two rounds is partial and different for different players. The fact that σ^{NoRev} is an equilibrium in all such alterations shows that σ is *information-proof*: no revelation of information (even if strategically coordinated by G-players and outsiders) could give any player an incentive to revise. Thus, structural robustness is substantially stronger than all the variants of the *ex post* Nash condition. (In the non-approximate notions, being *ex post* Nash is equivalent to being information proof. But in the approximate notions information proofness is substantially stronger.)

Remark 3 ***Invariance to order of play*** *G played sequentially*, $\mathcal{GS}$, is the following n-person extensive game. The n players are assigned types as in G. The play progresses sequentially, according to a fixed publicly known order. Every player, at his turn, knows all earlier choices.

Clearly, $\mathcal{GS}$ is an alteration of G, and every equilibrium σ of G has the following $\mathcal{GS}$ adaptation: At his turn, every player i chooses a pure-strategy with the same probability distribution σ_i as he does in the simultaneous-move game G. Structural robustness of σ implies that this adaptation of σ must be an equilibrium in every such $\mathcal{GS}$.

Moreover, the above reasoning continues to hold even if the order of play is determined dynamically, and even if it is strategically controlled by G-players and outsiders. Thus, a structurally robust equilibrium is invariant to the order of play in a strong sense.

Remark 4 ***Invariance to revelation and delegation*** *G with delegation*, $\mathcal{GD}$, is the following $(n + 1)$-players game. The original n G-players are assigned types as in G. In a first round of simultaneous play, every G-player chooses between (1) self-play and (2) delegate-the-play and report a type to an outsider, player $n + 1$. In a second round of simultaneous play all the self-players choose their own G pure strategies, and the outsider chooses a profile of G pure strategies for all the delegators. The payoffs of the G players are as in G; the outsider may be assigned any payoffs.

Clearly, $\mathcal{GD}$ is an alteration of G, and every equilibrium σ of G has adaptations that involve no delegation.

In the computer choice game, for example, consider an outsider with incentives to coordinate: his payoff equals one when he chooses the same computer for all delegators, zero otherwise. This alteration has a new (more efficient) equilibrium, not available in G: everybody delegates and the outsider chooses the most-reported type.

Nevertheless, as structural robustness implies, FC remains an equilibrium in $\mathcal{GD}$ (nobody delegates in the first round and they choose their favorite computers in the second). Moreover, FC remains an equilibrium under any scheme that involves reporting and voluntary delegation of choices.

Remark 5 ***Partially specified games*** Structurally robust equilibria survive under significantly more complex alterations than the ones above. For example, one could have multiple opportunities to revise, to delegate, to affect the order of play, to communicate, and more. Because of these strong invariance properties, such equilibria may be used in games which are only partially specified as illustrated by the following example.

Example 3 ***A game played on the Web*** Suppose that instead of being played in one simultaneous move, the Computer Choice game has the following instruction: 'Go to Web site xyz before the end of the week, and click in your computer choice.' This instruction involves substantial structural uncertainty: In what order would the players choose? Who can observe whom? Who can talk to whom? Can players sign binding agreements? Can players revise their choices? Can players delegate their choices? And so forth.

Because it is unaffected by the answers to such qsts, a structurally robust equilibrium σ of the one-simultaneous-move game can be played on the Web in a variety of ways without losing the equilibrium property. For example, players may make their choices according to their σ_i probabilities prior to the beginning of the click-in period, then go to the Web and click in their realized choices at individually selected times.

Remark 6 ***Competitive prices in Shapley–Shubik market games*** For a simple illustration, consider the following n-trader market game (see Shapley and Shubik 1977, and later references in Dubey and Geanakoplos 2003, and McLean et al. 2005). There are two fruits, apples and bananas, and a finite number of trader types. A type describes the fruit a player owns and the fruit he likes to consume. The players' types are determined according to individual independent prior probability distributions. Each trader knows his own type, and his payoff depends on his own type and the fruit he ends up with, as well as on the distribution of types and fruit ownership of his opponents (externalities are allowed, for example, a player may wish to own the fruit that most opponents like). In one simultaneous move, every player has to choose between (1) keeping his fruit and (2) trading it for the other kind.

The banana/apple price is determined proportionately (with one apple and banana added in to avoid division by zero). For example, if 199 bananas and 99 apples are traded, the price of bananas to apples would be $(199 + 1)/(99 + 1) = 2$, that is, every traded apple brings back two bananas and every traded banana brings back 0.5 apples.

With a small number of traders, the price is unlikely to be competitive. If players are allowed to re-trade after the realized price becomes known, they would, and a new price would emerge.

However, when n is large, approximate structural robustness implies being approximately information-proof. So even when the realized price becomes known, no player has significant incentive to re-trade, that is, the price is approximately competitive (Prob[some player can ε-improve his expected payoff by re-trading at the observed price] $\leq \rho$).

This is stronger than classical results relating Nash equilibrium to Walras equilibrium (for example, Dubey et al. 1980). First, being conducted under incomplete information, the above relates Bayesian equilibria to rational expectations equilibria (rather than Walras). Also the competitive property described here is substantially stronger, due to the immunity of the equilibria to alterations represented by extensive games. If allowance is made for spot markets, coordinating institutions, trade on the Web, and so on, the Nash-equilibrium prices of the simple simultaneous-move game are sustained through the intermediary steps that may come up under such possibilities.

Remark 7 ***Embedding a game in bigger worlds*** Alterations allow the inclusion of outside players who are not from G. Moreover, the restrictions imposed on the strategies and payoffs of the outsiders are quite limited. This means that alterations may describe bigger worlds in which G is embedded. Structural robustness of an equilibrium means that the small-world (G) equilibrium remains an equilibrium even when the game is embedded in such bigger worlds.

Remark 8 ***Self-purification*** Schmeidler (1973) shows that in a normal-form game with a continuum of anonymous players, every strategy can be purified, that is, for every mixed-strategy equilibrium one can construct a pure-strategy equilibrium (Ali Khan and Sun 2002 survey some of the large follow-up literature).

The *ex post* Nash property above constitutes a stronger (but asymptotic) result. Since the resulting play of a mixed strategy equilibrium yields pure-strategy profiles that are Nash equilibria (of the perfect information game), one does not need to construct pure-strategy equilibria: simply playing a mixed-strategy equilibrium yields pure-strategy profiles that are equilibria.

The approximate statement is: for every (ε, ρ) for sufficiently large n, Prob [ending at a pure strategy profile that is not an ε Nash equilibrium of the realized perfect information game] $\leq \rho$. Since both ε and ρ can be made arbitrarily small, this is asymptotic purification. Note that the model of Schmeidler, with a continuum of players, requires non-standard techniques to describe a continuum of independent random variables (the mixed strategies of the players). The asymptotic result stated here, dealing always with finitely many players, does not require any non-standard techniques.

Remark 9 ***'As if' learning*** Kalai and Lehrer (1993) show that in playing an equilibrium of a Bayesian repeated game, after a sufficiently long

time the players best-respond as if they know their opponents' realized types and, hence, their mixed strategies.

But being information-proof, at a structurally robust equilibrium (even of a one shot game) players' best respond (immediately) as if they know their opponents' realized types, their mixed strategies and even the pure-strategies they end up with.

Sufficient Conditions for Structural Robustness

Theorem 1 *Structural Robustness* (rough statement): the equilibria of large one-simultaneous-move Bayesian games are (approximately) structurally robust if

(a) the players' types are drawn independently, and
(b) payoff functions are anonymous and continuous.

Payoff anonymity means that in addition to his own type and pure-strategy, every player's payoff may depend only on aggregate data of the opponents' types and pure-strategies. For example, in the computer choice game a player's payoff may depend on her own type and choice, and on the *proportions* of opponents in the four groups: *I*-types who chose *I*, *I*-types who chose *M*, *M*-types who chose *I*, and *M*-types who chose *M*.

The players in the games above are only semi-anonymous, because there are no additional symmetry or anonymity restrictions other than the restriction above. In particular, players may have different individual payoff functions and different prior probabilities (publicly known).

The continuity condition relates games of different sizes and rules out games of the type below.

Example 4 ***Match the expert*** Each of n players has to choose one of two computers, *I* or *M*. Player 1 is equally likely to be one of two types: 'an expert who is informed that *I* is better' (*I*-better) or 'an expert who is informed that *M* is better' (*M*-better). Players 2, ..., n are of one possible 'non-expert' type. Every player's payoff is one if he chooses the better computer, zero otherwise. (Stated anonymously: choosing computer *X* pays one, if the proportion of the *X-better* type is positive, zero otherwise.)

Consider the equilibrium where player 1 chooses the computer he was told was better and every other player chooses *I* or *M* with equal probabilities. This equilibrium fails to be *ex post* Nash (and hence, fails structural robustness), especially as n becomes large, because after the play approximately one-half of the players would want to revise their choices to match the observed choice of player 1. (With a small n there may be 'accidental *ex post* Nash', but it becomes extremely unlikely as n becomes large.)

This failure is due to discontinuity of the payoff functions. The proportions of *I*-better types and *M*-better types in this game must be either $(1/n, 0)$ or $(0, 1/n)$, because only one of the n players is to be one of these types. Yet, whatever n is, every player's payoff is drastically affected (from 0 to 1 or from 1 to 0) when we switch from $(1/n, 0)$ to $(0, 1/n)$ (keeping everything else the same).

As n becomes large, this change in the type proportions becomes arbitrarily small, yet it continues to have a drastic effect on players' payoffs. This violates a condition of uniform equicontinuity imposed simultaneously on all the payoff functions in the games with $n = 1, 2, ...$ players.

See Also

- ▶ Internet, Economics of the
- ▶ Large Economies
- ▶ Purification
- ▶ Rational Expectations

Bibliography

Aumann, R.J., and L.S. Shapley. 1974. *Values of non-atomic games*. Princeton: Princeton University Press.

Bergemann, D., and S. Morris. 2005. Robust mechanism design. *Econometrica* 73: 1771–1813.

Cremer, J., and R.P. McLean. 1985. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica* 53: 345–361.

Dubey, P., and J. Geanakoplos. 2003. From Nash to Walras via Shapley–Shubik. *Journal of Mathematical Economics* 39: 391–400.
Dubey, P., and M. Kaneko. 1984. Information patterns and Nash equilibria in extensive games: 1. *Mathematical Social Sciences* 8: 111–139.
Dubey, P., A. Mas-Colell, and M. Shubik. 1980. Efficiency properties of strategic market games: An axiomatic approach. *Journal of Economic Theory* 22: 339–362.
Green, J.R., and J.J. Laffont. 1987. Posterior implementability in a two-person decision problem. *Econometrica* 55: 69–94.
Kalai, E. 2004. Large robust games. *Econometrica* 72: 1631–1666.
Kalai, E. 2005. Partially-specified large games. *Lecture Notes in Computer Science* 3828: 3–13.
Kalai, E., and E. Lehrer. 1993. Rational learning leads to Nash equilibrium. *Econometrica* 61: 1019–1045.
Khan, A.M., and Y. Sun. 2002. Non-cooperative games with many players. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland.
McLean, R., J. Peck, and A. Postlewaite. 2005. On price-taking behavior in asymmetric information economies. In *Essays in dynamic general equilibrium: Festschrift for David Cass*, ed. A. Citanna, J. Donaldson, H. Polemarchakis, P. Siconolfi, and S. Spear. Berlin: Springer. Repr. in *Studies in Economic Theory* 20:129–142.
Rustichini, A., M.A. Satterthwaite, and S.R. Williams. 1994. Convergence to efficiency in a simple market with incomplete information. *Econometrica* 62: 1041–1064.
Schmeidler, D. 1973. Equilibrium points of nonatomic games. *Journal of Statistical Physics* 17: 295–300.
Shapley, L.S., and M. Shubik. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85: 937–968.
Wilson, R. 1987. Game-theoretic analyses of trading processes. In *Advances in economic theory: Fifth world congress*, ed. T. Bewley. Cambridge: Cambridge University Press.

Laspeyres, Ernst Louis Etienne (1834–1913)

W. Erwin Diewert

Keywords
Drobisch price index; Index numbers; Laspeyres index; Laspeyres, E. L. E

JEL Classifications
B31

Laspeyres was born at Halle, Germany, on 28 November 1834 and died on 4 August 1913 at Giessen, Germany.

From 1853 to 1857, he studied at the universities of Tübingen, Berlin, Göttingen and Halle. He received a law degree from the University of Halle in 1857. He studied at the University of Heidelberg from 1857 to 1859, and in 1860 he obtained his Ph. D. from Heidelberg for the thesis, 'The Correlation between Population Growth and Wages'.

From 1860 until 1864 he worked as a lecturer at Heidelberg, where he wrote a history of the economic views of the Dutch (1863). In the following ten years, he taught at four different universities: 1864 – Basel; 1866 – the Polytechnic at Riga; 1869 – Dorpat; 1873 – Karlsruhe. Finally, from 1874 to 1900, he taught at the Justus-Liebig University at Giessen.

Laspeyres' main contribution to economics was his development of the index number formula that bears his name. Let the price and quantity of commodity n in period t be p_n^t and q_n^t respectively for $n = 1, \ldots, N$ and $t = 0, 1, \ldots, T$. Then the Laspeyres price index of the N commodities for period t (relative to the base period 0) is defined as

$$P_L \equiv \sum_{n=1}^{N} p_n^t q_n^0 / \sum_{n=1}^{N} p_n^0 q_n^0.$$

Laspeyres wrote his classic paper (1871), which suggested the above formula partly as an outgrowth of his empirical work on measuring price movements in Germany and partly to criticize the index number formula of Drobisch (1871). Using the notation defined above, the Drobisch price index for period t is defined as

$$P_D \equiv \left(\sum_{n=1}^{N} p_n^t q_n^t / \sum_{n=1}^{n} q_n^t \right) / \left(\sum_{n=1}^{N} p_n^0 q_n^0 / \sum_{n=1}^{N} q_n^0 \right).$$

Laspeyres criticized this formula by showing that the index generally changed even if all prices remained constant (that is, P_D does not satisfy an identity test, to use modern terminology). An even more effective criticism of P_D is that it is not invariant to changes in the units of measurement (whereas P_L is invariant).

Laspeyres did not write any further papers on index number theory. He wrote papers on economic history, the history of economic thought and on topical economic issues of his time; see Rinne (1981).

Selected Works

1863. *Geschichte der volkswirtschaftlichen Anschauungen der Niederländer und ihrer Literatur zur Zeit der Republik.* Leipzig.

1871. Die Berechnung einer mittleren Waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik* 16: 296–315.

Bibliography

Drobisch, M.W. 1871. Über die Berechnung der Veränderungen der Waarenpreise und des Geldswerths. *Jahrbücher für Nationalökonomie und Statistik* 16: 143–156.

Rinne, H. 1981. Ernst Louis Etienne Laspeyres 1834–1913. *Jahrböcher für Nationalökonomie und Statistik* 196: 194–216.

Lassalle, Ferdinand (1825–1864)

Tom Bottomore

Keywords

Cooperative production; Engels F.; Iron law of wages; Lassalle F.; Malthus's theory of population; Marx K. H.; Socialism

JEL Classifications

B31

Born in Breslau, 13 April 1825; died in Geneva, 31 August 1864. The only son of a prosperous Jewish silk merchant, Lassalle studied philosophy and history at the University of Breslau and subsequently at the University of Berlin, where he encountered the radical ideas of the 'Young Hegelians' and of the French socialist thinkers. During the 1848 revolution he was associated with Marx and the *Neue Rheinische Zeitung*, and was arrested for his activities but acquitted by a jury in 1849. In the course of his short and turbulent life (which ended as a result of an absurd duel with the former fiancé of a woman he wished to marry), Lassalle became known primarily as a political and economic theorist, and as a leading figure in the radical and working-class movements, who organized in 1863 the first socialist party in Germany (the General Union of German Workers).

Lassalle's economic ideas were derived to a large extent from Marx, often without acknowledgement, but he diverged from the latter in important respects. As Bernstein (1891) observed: 'Lassalle was much more indebted to Marx than he admitted in his writings, but he was a disciple of Marx only in a restricted sense.' The main divergence can be summarized as the substitution of an evolutionary conception of the movement from capitalism to socialism for Marx's idea of a revolutionary transition. In his 'Workers' Programme' (1862) and his 'Open Letter' (1863), Lassalle advocated a course of political action for the working-class movement with two principal aims: first, the achievement of universal and equal suffrage; second, the development, with state aid, of workers' cooperatives that would lead to a gradual socialization of the economy. His reliance upon the action of the state (conceived in the manner of Hegel rather than Marx) was very great, and in the 'Open Letter' he adduced an 'iron law of wages', derived from classical political economy, to show that neither individually nor collectively could workers improve their conditions of life except by replacing the wage system with self-employment (cooperative production), for which the necessary capital must be provided by the state. It was in this context that Lassalle responded to Bismarck's invitation (11 May 1863) to express his views on 'working class conditions and problems' and subsequently had several meetings with him; a course of action which Engels (letter to Kautsky, 23 February 1891) later assessed harshly as a step towards allying the workers' movement with German nationalism and the monarchy.

Marx had a low opinion of Lassalle's abilities as an economist and political thinker, and in his *Critique of the Gotha Programme* (1875) on the occasion of the unification of the two existing German workers' parties (the Social Democratic Workers' Party and the General Union of German Workers) he strongly criticized the Lassallean ideas which were embodied in the draft programme; in particular, the erroneous restriction of ownership of the instruments of labour to the capitalist class, excluding landowners, and the confused notion of an 'iron law of wages', which is simply, Marx argued, 'the Malthusian theory of population'.

Selected Works

1919–20. *Gesammelte Reden und Schriften,* 12 vols, edited with an Introduction by E. Bernstein. Berlin: Paul Cassirer.

Bibliography

Bernstein, E. 1891. *Ferdinand Lassalle as a social reformer.* London: Swan Sonnenschein, 1893. Reprinted, New York: Greenwood Press, 1969.

Footman, D. 1946. *The primrose path: A life of Ferdinand Lassalle*. London: Cresset Press.

Latent Variables

Dennis J. Aigner

A cursory reading of recent textbooks on econometrics shows that historically the emphasis in our discipline has been placed on models that are without measurement error in the variables but instead have stochastic 'shocks' in the equations. To the extent that the topic of errors of measurement in variables (or latent variables) is treated, one will usually find that for a classical single-equation regression model, measurement error in the dependent variable, y, causes no particular problem because it can be subsumed within the equation's disturbance term. But when it comes to the matter of measurement errors in the independent variables, the argument will usually be made that consistent parameter estimation is unobtainable unless repeated observations on y are available at each data point, or strong a priori information can be employed. The presentation usually ends there, leaving us with the impression that the errors-in-variables 'problem' is bad enough in the classical regression model and surely must be worse in more complicated models.

But in fact this is not so. For example, in a simultaneous equations setting one may employ over-identifying restrictions that appear in the system in order to identify error variances associated with exogenous variables, and hence to obtain consistent parameter estimates (not always, to be sure, but at least *sometimes*). Moreover, *ceteris paribus,* the dynamics in an equation can also be helpful in parameter identification. Finally, restrictions on a model's covariance structure, which are commonplace in sociometric and psychometric modelling, may also serve to aid identification. These are the three main themes of research with which we will be concerned.

To begin, let each observation (y_i, x_i) in a random sample be generated by the stochastic relationships:

$$y_i = \eta_i + u_i, \tag{1}$$

$$x_i = \xi_i + v_i, \tag{2}$$

$$\eta_i = \alpha + \beta\xi_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3}$$

Equation (3) is the heart of the model, and we shall assume $E(\eta_i/\xi_i) = \alpha + \beta\xi_i$, so that $E(\varepsilon_i) = 0$ and $E(\xi_i\varepsilon_i) = 0$. Also we denote $E\left(\xi_i^2\right) = \sigma_{\varepsilon\varepsilon}$. Equations (1) and (2) involve the measurement errors and their properties are taken to be $E(u_i) = E(v_i) = 0, E\left(u_i^2\right) = \sigma_{uu}, E\left(v_i^2\right) = \sigma_{vv}$ and $E(u_iv_i) = 0$. Furthermore, we will assume that the measurement errors are each uncorrelated with ε_i and with the latent variables η_i and ξ_i. Inserting the expressions $\xi_i = x_i - v_i$ and $\eta_i = y_i - u_i$ into (3), we get:

$$y_i = \alpha + \beta x_i + w_i, \tag{4}$$

where $w_t = \beta_i + u_i - \beta v_i$. Now since $E(v_i \mid x_i) \neq 0$, we readily conclude that least squares methods will yield biased estimates of a and β.

By assuming that all random variables are normally distributed we eliminate any concern over estimation of the ξ_i's as 'nuisance' parameters. This is the so-called *structural* latent variables model, as contrasted to the *functional* model, wherein the ξ_i's are assumed to be fixed variates. Even so, under the normality assumption no consistent estimators of the primary parameters of interest exist. This can be seen easily by writing out the so-called 'covariance' equations that relate consistently estimable variances and covariances of the observables (y_i and x_i) to the underlying parameters of the model. Under the assumption of joint normality, these equations exhaust the available information and so provide necessary and sufficient conditions for identification. They are obtained by 'covarying' (4) with y_i and x_i respectively. Doing so, we obtain:

$$\begin{aligned} \sigma_{yy} &= \beta\sigma_{yx} + \sigma_{\varepsilon\varepsilon} + \sigma_{uu}, \\ \sigma_{yx} &= \beta\sigma_{xx} - \beta\sigma_{vv} \\ \sigma_{xx} &= \sigma_{\xi\xi} + \sigma_{vv}. \end{aligned} \tag{5}$$

Obviously there are but three equations (involving three consistently estimable quantities, σ_{yy}, σ_{xx} and σ_{yx}) and five parameters to be estimated. Even if we agree to give up any hope of disentangling the separate influences of ε_i and u_i (by defining, say, $\sigma^2 = \sigma_{\varepsilon\varepsilon} + \sigma_{uu}$) and recognize that the equation $\sigma_{xx} = \sigma_{\xi\xi} + \sigma_{vv}$ will always be used to identify $\sigma_{\xi\xi}$ alone, we are still left with two equations in three unknowns (β, σ^2 and σ_{vv}).

The initial theme in the literature develops from this point. One suggestion to achieve identification in (5) is to assume that we know something about σ_{vv} relative to σ^2, or to σ_{xx}. Suppose this *a priori* information is in the form $\lambda = \sigma_{vv}/\sigma^2$ Then we have $\sigma_{vv} = \lambda\sigma^2$ and

$$\begin{aligned} \sigma_{yy} &= \beta\sigma_{yx} + \sigma^2, \\ \sigma_{yx} &= \beta\sigma_{xx} - \beta\lambda\sigma^2 \\ \sigma_{xx} &= \sigma_{\xi\xi} + \sigma_{vv}. \end{aligned} \tag{5a}$$

From this it follows that β is a solution to:

$$\beta^2\lambda\sigma_{yx}\left(\lambda\sigma_{yy} - \sigma_{xx}\right) - \sigma_{yx} = 0, \tag{6}$$

and that

$$\sigma^2 = \sigma_{yy} - \beta\sigma_{yx}, \tag{7}$$

Clearly this is but one of several possible forms that prior information on the underlying covariance structure may take (Jöreskog 1970); a Bayesian treatment also suggests itself (Zellner 1971).

Suppose that instead of having such information to help to identify the parameters of the simple model (1)–(3), there exists a variate z_i, observable, with the properties that z_i is correlated with x_i but uncorrelated with w_i. That is, z_i is an *instrumental variable* (for x_i). This is tantamount to saying that there exists another equation relating z_i to x_i, for example,

$$x_i = \gamma z_i + \delta_i, \tag{8}$$

with $E(z_i\delta_i) = 0$, $E(\delta_i) = 0$ and $E\left(\delta_i^2\right) = \sigma_{\delta\delta}$, Treating (4) and (8) as our structure (multinormality is again assumed), and forming the covariance equations we get, in addition to (5):

$$\begin{aligned} \sigma_{yz} &= \beta\sigma_{yz}, \\ \sigma_{xx} &= \gamma\sigma_{zx} - \sigma_{\delta\delta}, \\ \sigma_{zx} &= \gamma\sigma_{zz}. \end{aligned} \tag{9}$$

It is apparent that the parameters of (8) are identified through the last two of these equations. If, as before, we treat $\sigma_{\varepsilon\varepsilon} + \sigma_{uu}$ as a single parameter, σ^2, then, (5) and the first equation of (9) will suffice to identify β,σ^2,σ_{vv} and $\sigma_{\xi\xi}$.

This simple example illustrates how additional equations containing the same latent variable may serve to achieve identification. This 'multiple equations' approach spawned the revival of latent variable models in the 1970s (Zellner 1970; Jöreskog and Goldberger 1975).

From consideration of (4) and (8) together we saw how the existence of an instrumental variable (equation) for an independent variable subject to measurement error could resolve the identification

problem posed. This is equivalent to suggesting that an over-identifying restriction exists somewhere in the system of equations from which (4) is extracted to provide an instrument for a variable like x_i. But over-identifying restrictions cannot be traded-off against measurement error variances without qualification. Indeed, the *locations* of the exogenous variable measured with error, relative to those of the over-identifying restrictions appearing elsewhere in the equation system, are crucial (Geraci 1976). To elaborate, consider the following equation system,

$$\begin{aligned} y_1 + \beta_{12} y_2 &= \gamma_{11}\xi_1 + \varepsilon_1, \\ \beta_{21} y_1 + y_2 &= \gamma_{22}\xi_2 + \gamma_{23}\xi_3 + \varepsilon_2, \end{aligned} \tag{10}$$

where ξ_j ($j = 1, 2, 3$) denote the latent exogenous variables in the system. Were they regarded as *observable,* the first equation – conditioned on this supposition – is over-identified (one over-identifying restriction), while the second equation is conditionally just-identified. Therefore, at most one measurement error variance can be identified.

Consider first the specifications $x_1 = \xi_1 + v_1$, $x_2 = \xi_2$, $x_3 = \xi_3$, and let σ_{11} denote the variance of v_1. The corresponding system of covariance equations, under our standard assumption of multinormality, suffices to examine the state of identification of all the parameters. There are six equations available to determine the six unknowns, β_{12}, β_{21}, γ_{11}, γ_{22}, γ_{23}, and σ_{11}, and in this case all parameters are exactly identified. Were the observation error instead associated with ξ_2, the conclusion would be different. Under that specification β_{12} and γ_{11} are overdetermined, while there are only three covariance equations available to solve for β_{21}, γ_{22}, γ_{23} and σ_{22}. Hence these latter four parameters, all of them associated with the second equation in (10), are not identified.

The results presented and discussed thus far apply only to models depicting *contemporaneous* behaviour. When dynamics are introduced into either the dependent or the independent variables in a linear model with measurement error, the results are usually beneficial. To illustrate, we revert to a single-equation setting, one that parallels the development of (4). In particular, suppose that the sample at hand is a set of time-series observations and that (4) is instead:

$$\begin{aligned} \eta_t &= \beta\eta_{t-1} + \varepsilon_t, \\ y_t &= \eta_t + u_t, \quad t = 1, \ldots, T, \end{aligned} \tag{11}$$

with all the appropriate previous assumptions imposed, except that now we will also use $|\beta| < 1$, $E(u_t) = E(u_{t-1}) = 0, E\left(u_t^2\right) = E\left(u_{t-1}^2\right) = \sigma_{uu}$, and $E\left(u_t u_{t-1}\right) = 0$. Then, analogously to (5), we have:

$$\begin{aligned} \sigma_{yy} &= \beta\sigma_{yy-1} + \sigma_{\varepsilon\varepsilon} + \sigma_{uu}, \\ \sigma_{yy-1} &= \beta\left(\sigma_{yy} - \sigma_{uu}\right), \end{aligned} \tag{12}$$

where σ_{yy-1} is our notation for the covariance between y_t and y_{t-1} and by assumption we equate the variances of y_t and y_{t-1}. This eliminates one parameter from consideration (σ_{vv}), and there is now a system of two equations in only three unknowns. Unfortunately, we are not now helped any further by our earlier agreement to combine the effects of the equation disturbance term (ε_t) and the measurement error in the dependent variable (u_t).

Fortunately, however, there is some additional information that can be utilized to resolve things: it lies in the covariance between current y_t and lags beyond one period (y_{t-s} for $s \geq 2$). These covariances are of the form:

$$\sigma_{yy-s} = \beta\sigma_{yy-s+1}, \quad s \geq 2, \tag{13}$$

so that any one of them taken in conjunction with (12) will suffice to solve for β, $\sigma_{\varepsilon\varepsilon}$, and σ_{uu}. See Maravall and Aigner (1977) and Hsiao (1977) for more details and extension to the simultaneous equations setting.

Structural modelling with latent variables is not only appropriate from a conceptual viewpoint, in many instances it also provides a means to enhance model specifications by taking advantage of information that otherwise might be misused or totally ignored. Several interesting applications of latent variables models in econometrics have appeared since the early 1970s. Numerous others have been published in the psychometrics and sociometrics literature over the years. An in-depth presentation of the theoretical developments outlined here with

references to the applied literature is contained in Aigner et al. (1984).

See Also

- ▶ Errors in Variables
- ▶ Instrumental Variables
- ▶ Principal Components

Bibliography

Aigner, D.J., Hsiao, C., Kapteyn, A. and T. Wansbeek. 1984. Latent variable models in econometrics. Chapter 23 in *Handbook of econometrics*, ed. Z. Griliches., and M. Intriligator, vol. 2. Amsterdam: North-Holland.

Geraci, V. 1976. Identification of simultaneous equations models with measurement error. *Journal of Econometrics* 4: 263–283.

Hsiao, C. 1977. Identification of a linear dynamic simultaneous error-shock model. *International Economic Review* 18: 181–194.

Jöreskog, K. 1970. A general method for the analysis of covariance structures. *Biometrika* 57: 239–251.

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: 631–639.

Maravall, A. and D.J. Aigner. 1977. Identification of the dynamic shock-error model: The case of dynamic regression. Chapter 18 in *Latent variables in socioeconomic models,* ed. D.J. Aigner., and A.S. Goldberger. Amsterdam: North-Holland.

Zellner, A. 1970. Estimation of regression relationships containing unobservable independent variables. *International Economic Review* 11: 441–454.

Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.

Latifundia

E. V. K. FitzGerald

The *latifundium* first appears extensively during the later Roman empire as a type of large agricultural enterprise which obtained labour services from a resident workforce (*coloni*) in return for the temporary use of a plot of land, when the slaves on the estates created from the land of conquered communities became too costly. As the Empire declined, the latifundia also became local centres of economic and political power, absorbing the free peasantry into villein or 'servile' status, and providing the foundation for the manorial system of rural organization in the Middle Ages (Tuma 1965). Labour shortages and urban growth led to an abandonment of this form of direct exploitation of labour in Western Europe by the 15th century in favour of more flexible rental agreements in kind, and eventually in money; though serfdom persisted in Eastern Europe and Russia well into the 19th century, and became a central theme in the 'agrarian question' (Hussain and Tribe 1981).

The reconquest of Spain in the 15th century had confirmed the latifundium as an effective means of territorial and labour control based on large land grants to military leaders, so the system was logically extended to Latin America, where the *hacienda* (or large autonomous landed estate) became the cornerstone of colonial policy (Florescano 1984). In the 17th and 18th centuries, hacienda autonomy was strengthened by the weakness of colonial administration from the cities. After Independence in the early 19th century, the agrarian structure of Latin America was perpetuated for over a hundred years by the central economic role of raw material exports produced by large landowners, and the lack of access to political power of the peasantry on sub-subsistence plots, known as *minifundio* (Barraclough 1973).

Detailed historical research (reported in Duncan and Rutledge 1977) reveals the enormous variety of *latifundio* arrangements for securing labour by ceding subsistence plots, and conflicts with neighbouring Indian communities over such land rights. It has also shown that these enterprises were generally market-oriented and guided by a profit motive, although there are close parallels with the manorial system (Kay 1974). Nonetheless, the archaic and feudal nature of the latifundio has traditionally been seen (from Mariategui 1928 to Furtado 1970) as an obstacle to economic development, engendering a theoretical debate as to whether Latin American agriculture should be

seen as capitalist because of its mercantile relationship with the national and world economy (Frank 1967) or as feudal because of its labour relations (Laclau 1971).

In political doctrine, the latter view has tended to prevail, and latifundia have been the main target of land reform in Latin America as involving inequitable social relations and inefficient use of land. Occupying up to half the farmed area as recently as World War II, they are now almost extinct. The latifundio was not a feature of the rest of the Third World (except the Philippines) where large estates are usually organized as plantations or on a sharecropping basis, although in ancient and colonial times various forms of obligatory labour contributions from the peasant communities were common.

In underdeveloped areas, income cannot be realized from land without intensive labour use, so any pattern of distribution of property rights is necessarily accompanied by a system of interpersonal and intergroup relationships governing the application of labour to land. The historical survival of the 'servile' system in various forms has lead to a reassessment of its economic logic, particularly in comparison to the large 'commercial' farm employing exclusively wage labour (de Janvry 1981). The cost of such servile labour power is less than the price of proletarian labour power (i.e. the free market wage) because the opportunity cost to the owner of the ceded plot is less than the value of production the labourer can generate on it through the use of family labour. The extensive use of land on the latifundio means a low opportunity cost of marginal plots and its denial to independent smallholders; while the lack of alternative occupations means that the labour power of *colono*'s family has a near-zero opportunity cost as well. To be effective, the local hacienda system must thus prevent outward migration by mechanisms of a legal or traditional nature: it is characteristic of a situation where there is a scarcity of rural labour and landlord dominance of local society. The system should also be seen in the context of widespread reciprocal labour agreements and payments in labour time (for rent, use of draught animals, etc.) between peasant farmers themselves (Pearce 1975). From the landowner's point of view, therefore, the latifundio system can be a profit-maximizing solution, and in a situation where labour rather than land (despite appearances) is really the scarce resource, it may be a relatively technically efficient (albeit not socially desirable) solution for a capitalist economy as a whole.

In a situation of relative labour surplus (commonly associated with early industrialization, population transition and the modernization of agriculture itself), the 'semi-proletarian' settled outside the estate becomes an even cheaper and more flexible source of labour power, as labour is only used and paid seasonally, while the minifundio, now producing for subsistence and even some marketed surplus, can deliver cheap labour without reciprocal obligations to the capitalist sector, made up now of latifundia in transition towards commercial farms with mechanization and technical inputs (Goodman and Redclift 1981).

Although the latifundio as such is becoming a thing of the past, similar systems of rural labour organization persist because the nature of agricultural production itself is such that the need for seasonal labour for large scale-efficient farms coexists with the superior work-intensity of household production. This implies that the articulation of distinct forms of production in a single location is still necessary. The equivalent of the latifundio concept in a post-capitalist context might be detected in the form of state farms or producer collectives established in Eastern Europe, where household labour time is divided between collective enterprise land and the family plot. This ensures that necessary labour supply for harvests etc. on the mechanized collective land, while providing an income incentive for high productivity on labour-intensive individual land. This system need not necessarily be exploitative (in that the profits so generated are not appropriated by a landowner) but apparently must still be maintained by non-market mechanisms such as collective solidarity or restraints on migration.

See Also

► Land Reform
► Peasant Economy

Bibliography

Barraclough, S. 1973. *Agrarian structure in Latin America*. Lexington: Heath.
Delgado, O. (ed.). 1965. *Reforma agraria en America Latina*. Mexico City: Fondo de Cultura Economica.
de Janvry, A. 1981. *The agrarian question and reformism in Latin America*. Baltimore: Johns Hopkins Press.
Duncan, K., and I. Rutledge (eds.). 1977. *Land and labour in Latin America: Essays on the development of agrarian capitalism in the nineteenth and twentieth century.* Cambridge: Cambridge University Press.
Florescano, E. 1984. The formation and economic structure of the hacienda in New Spain. In *The Cambridge history of Latin America, vol II: Colonial Latin America*, ed. L. Bethell. Cambridge: Cambridge University Press.
Frank, A.G. 1967. *Capitalism and underdevelopment in Latin America*. New York: Monthly Review Press.
Furtado, C. 1970. *Economic development of Latin America*. Cambridge: Cambridge University Press.
Goodman, D., and M. Redclift. 1981. *From peasant to proletarian: Capitalist development and agrarian transitions*. Oxford: Blackwell.
Griffin, K. 1981. *Land concentration and rural poverty*, 2nd ed. London: Macmillan.
Hussain, A., and K. Tribe. 1981. *Marxism and the agrarian question*. London: Macmillan.
Kay, C. 1974. Comparative development of the European manorial system and the Latin American hacienda system. *Journal of Peasant Studies* 2(1): 69–98.
Laclau, E. 1971. Feudalism and capitalism in Latin America. *New Left Review* 67: 19–38.
Mariategui, J.C. 1928. *Siete ensayos de interpretacion de la realidad peruana*. Lima: Editorial Amauta.
Pearce, A. 1975. *The Latin American peasant*. London: Cass.
Tuma, E.H. 1965. *Twenty-six centuries of agrarian reform*. Berkeley: University of California Press.

Latin American Economic Development

Mauricio Cárdenas and Steven M. Helfand

Abstract

This article examines the strategies, successes and failures of economic development in Latin America since 1870. We divide the analysis into four key development phases: primary export-led growth (1870–1929), import substitution industrialisation (1945–1982), debt crisis (1980s) and the Washington Consensus (1990s). We demonstrate progress on many fronts, but underscore two key challenges for the region. One of them relates to weak institutions and state capacity; the other is the persistence of high levels of poverty and inequality. We conclude with a discussion of these challenges and of specific actions that are necessary to accelerate development in the region.

Keywords

Latin America; Economic development; Primary export-led growth; Import substitution industrialization (ISI); Debt crisis; Washington consensus; Institutions; State capacity; Poverty; Inequality

JEL Classification

N16; O1; O2; O54

Introduction

The countries in the Western Hemisphere that are former colonies of Spain and Portugal not only share a common heritage but also similar patterns of development. After a disastrous period of economic performance in the decades following independence in the early nineteenth century, Latin American countries pursued a primary export-led model of development at the turn of the twentieth century and then implemented an inward looking industrialisation strategy following the Second World War. After struggling with a severe crisis in the 1980s, the region embraced the market-driven Washington Consensus in the early 1990s. The current phase of development is being shaped by a renewed recognition of the importance of the state, increased attention to the high levels of poverty and inequality, and the emergence of China as a leading force in global trade. Concerns associated with the primary export-led model of development are again at the top of the agenda.

Common endowments and institutions explain remarkable similarities in terms of living standards, commodity export dependence and high levels of economic inequality. GDP per capita in PPP terms, for example, only varied by a factor of 5.4 between Nicaragua (the lowest) and Chile (the highest) in 2007, in comparison to Asia, where it varied by a factor of 32 between Nepal and Japan (Milanovic 2011). There are also important differences in the region. Size is a case in point: the seven largest countries in the region account for over 80 % of the population and GDP. Brazil, alone, has around one third of the region's population and GDP. Geographical location is another differentiating factor. Mexico and the Central American and Caribbean countries are more economically dependent on the USA than countries in South America. This is in part due to trade and tourism, but also to labour migration and remittances. Ethnic differences are a third factor that have impacted economic and social outcomes. Slavery was much more important in the Caribbean and the northeast of Brazil, while late nineteenth and early twentieth century European immigration played a much more significant role in the Southern Cone countries. It is the similarities that make a study of Latin America relevant, and it is the differences that force analysts to search beneath the generalisations in order to understand the region's heterogeneity.

Between 1870 and 1981, real GDP per capita increased by nearly eightfold in the region, which is faster than growth in any other region of the world, and comparable to growth in the USA (Table 1). However, during the subsequent 20 years, Latin America was among the slowest growing regions in the world. In the years immediately before and after the global recession of 2008–2009, Latin America once again experienced relatively fast growth, in part as a consequence of better macroeconomic policies, but also as a result of the economic tailwinds from China. A wave of optimism spread throughout the region in the new millennium as living standards increased at a rate not experienced since the 1970s.

Two of the most daunting challenges that Latin America faces relate to the low quality and effectiveness of institutions, and the persistence of poverty and inequality. There is an increasing awareness that the quality of both state and market institutions needs to improve significantly. The Latin American state has many important functions that need to be carried out more effectively. Well-functioning markets also rely on institutions – legal, anti-trust, regulatory – which are still far below the standards of the developed world. It is also the case that Latin American countries remain among the most unequal in the world. Yet inequality has been falling for more than a decade in many countries, and the rate of poverty reduction has accelerated. Although there are grounds for optimism based on recent performance, Latin America's future hinges crucially on continued progress in these two areas.

Primary Export-Led Growth: 1870–1929

Most Latin American countries gained independence in the first quarter of the nineteenth century. The subsequent 50 years were characterised by political instability and slow growth. Although data are scarce, between 1820 and 1870 income per capita appears to have been stagnant in the region (Madison 2008). In contrast, between 1870 and 1913, Latin America pursued a primary export model of development and grew faster than any other region in the world. Real income per capita rose at an annual rate of over 1.8 % for the eight Latin America countries shown in Table 1. Growth in Argentina was particularly impressive in this period, with 1913 income per capita surpassing the Western European average. Latin American growth continued at 1.5 % per year from 1913 to 1929.

Expansion was largely extensive, based on the incorporation of new land and additional labour. In some countries, especially in the Southern Cone, immigrants also contributed to population growth and an increase in human capital. The export boom was accompanied by capital inflow that helped finance investments in infrastructure. The length of railroad tracks in the region, for example, increased by a factor of 12 between 1870 and 1900, from about 50,000 to 60,000 km (Thorp 1998).

Latin American Economic Development, Table 1 Real GDP per capita by region

	1870	1870–1913	1913–1929	1929–1945	1945–1972	1972–1981	1981–1990	1990–00	2000–2008	1870–1981	1981–2008	2008
Region	1990 $	Average Annual Growth by Period										1990 $
Western Europe (12)	2080	1.34	1.10	− 0.35	3.87	2.14	2.05	1.76	1.34	1.73	1.73	22,246
United States	2445	1.82	1.66	3.36	1.15	1.88	2.33	2.07	1.14	1.86	1.88	31,178
Former USSR	943	1.06	− 0.44	–	–	1.47	0.77	− 4.26	7.42	1.74	0.77	7904
Latin America (8)	742	1.83	1.50	0.72	2.61	2.52	− 0.63	1.63	2.16	1.87	1.02	7614
Latin America	676	1.86	–	–	–	2.51	− 0.62	1.52	2.13	1.88	0.98	6973
Asia (16)	546	0.50	–	–	–	2.96	3.86	3.37	5.25	1.14	4.09	5673
Africa	500	0.57	–	–	–	0.87	− 0.46	0.16	2.62	0.99	0.67	1780
World Average	870	1.31	–	–	–	1.65	1.45	1.60	2.94	1.50	1.95	7614

Notes: Based on 1990 International Geary-Khamis dollars
Western Europe (12) includes Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Norway, Sweden, Switzerland, and the U.K
Latin America (8) includes Argentina, Brazil, Chile, Colombia, Mexico, Peru, Uruguay, and Venezuela
Latin America includes Latin American (8), 15 others, and 21 Caribbean countries
Asia (16) includes China, India, Pakistan, Bangladesh, Indonesia, Japan, Philippines, South Korea, Thailand, Taiwan, Burma, Hong Kong, Malaysia, Nepal, Singapore, and Sri Lanka
Source: Data download from Angus Madison, 5/2011, http://www.ggdc.net/MADDISON/oriindex.htm

Exports were extremely concentrated in a small number of products, creating a vulnerability to commodity booms and busts that could last for decades. Brazil, El Salvador, Guatemala, Haiti and Nicaragua, for example, earned between 62 and 85 % of export earnings from coffee alone around 1913, while Bolivia, Chile, Cuba, Ecuador and Panama earned at least 64 % from a single product (Bulmer-Thomas 2003).

The benefits for long-term growth derived from building institutions, infrastructure, and a local market varied by product and country size. In terms of products, locations where factor endowments had been favorable to products with economies of scale (such as sugar) developed more extractive institutions that contributed to the persistence of high economic inequality (Sokoloff and Engerman 1997). In large countries, exporters had greater incentives to invest their profits in activities geared toward the domestic market, thus stimulating the growth of manufactured products (as in Colombia and Brazil). But the size of the country and the product that predominated were not the only factors that mattered. Differences in the composition of elites and the resultant patterns of land occupation help to explain the much more favorable twentieth century outcomes in Colombia and Costa Rica relative to El Salvador and Guatemala – all countries in which coffee was extremely important (Nugent and Robinson 2010).

The global disruptions that took place between 1914 and 1945 threatened the sustainability of the primary export-led model of growth. The depression of the 1930s reduced demand and prices for Latin American exports. The two world wars interrupted trade routes with Europe and made manufactured products difficult to import. These events created incentives and opportunities to industrialise. Large countries, and those with more autonomous public sectors (such as Uruguay and Costa Rica), took advantage of these opportunities in the 1930s. They abandoned the gold standard earlier and experimented with trade, credit and other policies that actively encouraged local manufacturing. Small countries, and those with more dependent governments (such as Cuba), remained on the gold standard longer and responded much more passively to the new external environment (Diaz-Alejandro 1984). The experiences of state-led industrialisation in the USSR and the New Deal in the USA provided examples of a much more active economic role of the state and became important references for governments in the region. Thus, the period between 1929 and 1945 was one of transition between different models of development. But it would only be after the Second World War that an inward looking model of industrialisation would emerge as the predominant strategy of development in the region (Baer 1972).

Import Substitution Industrialisation: 1945–1982

Import substitution industrialisation (ISI) emerged as a consequence of the disruption in global trade between 1929 and 1945 and at a time when there was growing dissatisfaction with the primary export-led model. An influential group of Latin American 'structuralist' economists led by Raul Prebisch at the UN Economic Commission for Latin America (ECLA) argued that the primary export-led model was unable to provide sustained improvements in living standards. Two key tenets of the structuralist school were the centre–periphery model of the world economy and a hypothesis regarding the secular decline in the terms of trade of primary exporters (Prebisch 1950). Latin American countries, as well as other 'peripheral' countries, depended on the centre not just as a market for exports of primary products, but also as a source of capital and technology. The periphery's form of insertion into the world economy had a number of negative consequences for their socioeconomic structures, including slow productivity growth, low wages and small domestic markets. The decline in the terms of trade, resulting from the slow growth in the demand for primary products and less competitive markets for industrial products in the centre countries, was regarded as an additional obstacle to development.

According to the structuralist school, economic development in Latin America required

industrialisation, which involved protecting the domestic market from external competition and an active role of the state in promoting strategic sectors. At roughly the same time as the ECLA economists were writing about ISI, advances in economic theory and the emergence of 'development economics' provided many of the economic arguments that justified state interventions to deal with market imperfections. Latin American intellectuals continued to contribute to theories of development through the dependency school that flourished in the 1960s and 1970s (Kay 1989). The reformist branch of the dependency school (e.g., Cardoso and Faletto 1969) criticised ISI, but believed, like the ECLA economists, that capitalist industrial development was possible in Latin America. The radical branch sought to develop a Marxist theory of dependency and believed that a socialist revolution was necessary (e.g., Dos Santos 1970).

The ISI policy toolbox included trade protection through tariff and non-tariff instruments, multiple exchange rates that were typically lower for imports of capital and intermediate goods, active industrial policy, and supportive fiscal and monetary policies (Franko 2007). The early stages of ISI focused on creating industries to produce basic consumer and durable goods. Tariffs and import quotas increased the price of these goods and restricted their quantity, thus increasing the incentives for local production. By 1960, nominal protection rates on durable goods were over 90 % in Chile, Colombia and Mexico, and over 250 % in Argentina and Brazil (Bulmer-Thomas 2003). Overvalued currencies lowered the price of the imported capital goods that were necessary for industrialisation. Tax breaks and subsidies further increased incentives for domestic production, and national development banks, such as CORFO in Chile and BNDES in Brazil, provided equity and long-term credit for key industrial projects.

The growth of local industry increased the demand for inputs, such as metals, chemicals and electricity, as well as for transportation and other crucial services. Because of the scale of these projects, the time horizon necessary to recoup the investments and shallow domestic capital markets, governments often undertook these projects through state-owned enterprises (SOEs). By the 1970s and 1980s there were more than 500 SOEs in Chile and Brazil, and over 1000 in Mexico (Edwards 1995). In many cases the SOEs dominated the sectors in which they operated. In Brazil, for example, over 95 % of assets in railways, ports, telegraph and telephone, and water, gas and sewers were held by SOEs. In chemicals, mining and electricity, the share was over 50 % (Evans 1979). The domestic private sector in many Latin American countries also found it difficult to move beyond consumer durables and compete in the production of capital goods or in sectors that required more advanced technology. Thus, many of the most dynamic sectors were dominated by multinational corporations (MNCs). Around 1970, for example, in Chile, Colombia, Mexico and Peru, the foreign share of domestic production in chemicals, transport equipment, and electrical machinery varied between 49 and 80 % (Jenkins 1984).

Criticism of ISI's shortcomings began to emerge from many quarters in the 1960s (see Hirschman 1968). Yet it was only in retrospect, as Hirschman (1987) emphasises, that the achievements of this period – characterised by ISI, increased urbanisation and greater labour market participation – could be fully appreciated. Although the 1950s, 1960s and 1970s, were the decades with the highest rates of population growth in the twentieth century, real income per capita rose by 2.6 % per year between 1945 and 1981 in the Latin American 8 (Table 1). These 36 years compare quite favourably with the first 45 years of the century when income per capita rose by 1.4 % per year, and with the subsequent 27 years shown in Table 1 when income per capita rose by only 1.0 % per year. Total lifetime income rose by much more than income per capita in this period, as life expectancy rose from 40 years in 1940 to 65 years in 1980. (These data are from Thorp (1998) and are based on Argentina, Brazil, Chile, Colombia, Mexico and Venezuela. Data from these six countries are almost identical to the broader set of 20 countries for which data are available for a more limited period of time. The data on illiteracy in this paragraph refer to 20 countries and also come from Thorp (1998))

Improvements in education, as measured by the illiteracy rate, tell a similar story. Illiteracy among the population age 15 and over fell from 49 to 21 % between 1940 and 1980. There is no question that the middle class expanded and living standards improved dramatically for a large portion of the population in the region.

The ISI model in Latin America did, nevertheless, have many shortcomings. First, it generated considerable inefficiency. The ISI strategy was supposed to protect infant industries for a limited period of time while domestic companies learned to work with new technologies, increased their scale and lowered production costs. As a result of rent seeking, however, protection and subsidies often continued indefinitely. Second, unlike with East Asian countries such as Japan and South Korea, the transition from ISI to an export-oriented development strategy was not successful in Latin America. This was particularly harmful to the smaller countries with little opportunity to achieve economies of scale. The failure to move past ISI contributed to slower growth in total factor productivity relative to the Asian Tigers, and this had long-term consequences for GDP growth and living standards. Third, the policies that were intended to subsidise industry implicitly taxed agriculture, thus exacerbating rural poverty and contributing to excessively rapid urbanisation. Fourth, by reducing the price of imported capital goods, the development model of this period favoured capital-intensive sectors that were unable to create enough employment to keep pace with a rapidly increasing urban labour force. The result was high informality, with negative consequences for the distribution of income.

The Debt Crisis and the Lost Decade of the 1980s

One of the principal weaknesses of ISI in the 1950s and 1960s was an insufficiency of export earnings and, thus, the pervasiveness of current account deficits which resulted in frequent balance of payments crises. The perennial shortage of foreign exchange was temporarily eased in the 1970s as a result of the large current account surpluses in oil exporting countries after the first oil price shock in 1973. While country experiences differed, rather than fully adjust to the new international terms of trade, most Latin American governments continued to target high growth rates by taking on massive amounts of debt from private international banks at low, yet adjustable, interest rates (Fishlow 1986). Between 1970 and 1982, medium-and long-term debt in Mexico rose from US$7 billion to US$88 billion. In Brazil and Argentina, debt rose respectively from US$5 billion to US$83 billion, and from US $4 billion to US$47 billion (Sachs 1990).

The real difficulties came as a result of the second oil shock in 1979 which again drove OECD countries into recession and reduced the demand for Latin American exports. As part of the fight against stagflation in the USA, interest rates were raised above 8 % in real terms in 1981. Thus, the attractiveness of negative real interest rates in the 1970s quickly became a liability. Mexico declared a moratorium on its debt payments the next year, and the crisis quickly spread throughout the region. Latin American countries that had become accustomed to net resource inflows of around US$10 billion per year in the late 1970s suddenly had to generate net outflows in excess of US$30 billion by 1983 (Edwards 1995). Thus, while unsustainable policies had created vulnerabilities, the external shocks triggered the crisis.

What began as a debt problem in 1982 eventually became a much broader crisis that extended into the 1990s. Real GDP per capita fell 3 years in a row starting in 1981, and repeated the terrible performance in the years 1988–1990. Between 1981 and 1990, real income per capita fell at an annual rate of –0.62 % (Table 1). Some countries did not recover their pre-crisis levels of income per capita until the mid-1990s, thus giving rise to the term the ‘lost decade’ in Latin America. Average inflation was close to 100 % per year between 1982 and 1986, rising to over 200 % per year between 1987 and 1992 (Edwards 1995). In 1990, four countries had inflation rates over 1000 %. As a result of slow growth and contracting government expenditures, poverty increased sharply in the 1980s. The number of

people living on less than two dollars a day rose from 91 to 131 million between 1980 and 1989 (Morley 1995).

International lenders initially diagnosed the problem as one of liquidity, not solvency, and thus were unwilling to forgive any significant amount of debt, especially for the larger countries. Orthodox stabilisation plans based on demand repression were initially adopted throughout the region in order to address the main symptoms of the crisis – fiscal and balance of payments deficits, and inflation. When these plans failed to control inflation, some countries experimented with heterodox plans that incorporated wage and price freezes, and introduced new currencies, as in the cases of the Austral plan in Argentina and the Cruzado plan in Brazil. Structural adjustment policies eventually began to complement, or replace, stabilisation plans in the 1980s as the focus shifted from demand management to supply.

While a number of creative solutions were attempted to reduce the debt burden, including debt-for-equity and debt-for-nature swaps, and making use of a secondary market that had emerged to trade discounted sovereign debt, it was not until 1989 that US Secretary of Treasury Brady promoted a plan that would eventually restructure Latin America's debt. There was a menu of options that countries could choose from – including extended maturity dates, reduced face value of loans, and lower interest rates – that were all intended to reduce the burden of debt servicing. Multilateral institutions provided loan guarantees on the new debt in order to induce private banks to participate. The Brady plan, and many of the Washington Consensus reforms that will be discussed in the next section, contributed to restarting growth. In the final analysis, it was only through growth that the debt crisis faded.

The Washington Consensus of the 1990s

The need for a new development paradigm was a natural consequence of the discontent with the inward-looking growth strategy of the post-WWII period and the disastrous economic outcomes of the 1980s. Latin America was emerging from a severe episode of debt-overhang in the early 1990s, and consequently began to emphasise macroeconomic stability as a prerequisite for growth. At the same time, a shift towards a market-based development approach was favoured by the International Financial Institutions.

Chile was the first country in the region to transition to an export-oriented, market-based strategy. The transformation began in the mid-1970s, but was not accompanied by macroeconomic stability in its first decade. The country experienced several deep recessions and, for somewhat different reasons, also suffered from the 1980s debt crisis that hit the rest of the region. Poverty rose sharply in this period, and the liberalising strategy was tainted by having been imposed by an authoritarian regime. Yet important changes took place in this period that contributed to future success. These included the growth and diversification of exports, improvements in government budgeting, and modernisation of the business community (Ffrench-Davis 2002). Chile only became a model for many other countries in the region as of the 1990s. This coincided with the take-off of its economy, a return to democracy, and a more explicit policy focus on growth with equity. Income per capita grew by 4.9 % per year in the 1990s (Madison 2008), or three times the Latin American average, and poverty fell by half (ECLAC 2004). In addition to the successes of the 1990s, Chile began to reform the reforms about a decade before most other countries. This contributed to its role as a policy leader.

Summarising the convergence of views between officials in Washington and technocrats in Latin America in the late 1980s, Williamson (1990) listed a Decalogue of policies that became known as the Washington Consensus. Trade and foreign direct investment liberalisation, the so-called *apertura*, as well as the deregulation of key markets and the privatisation of state-owned enterprises were its core components. Fiercely challenged in the political and ideological arena, the Washington Consensus became an expression synonymous with 'neo-liberalism'.

The new paradigm also underscored the importance of policies that resulted in fiscal discipline,

market-determined interest rates and competitive exchange rates. Although not directly stating the means to achieve these ends, the Washington Consensus was soon associated with the need to provide independence to central banks. This was seen as a prerequisite to curb inflationary financing of the fiscal deficit and the use of preferential interest rates for specific groups. Budget deficits and overly rigid exchange rates, however, continued to plague many countries and would contribute to a new round of crises at the end of the 1990s.

Trade and capital account liberalisation, financial deregulation, privatisation and central bank independence, were widely adopted in the region during the late 1980s and early 1990s through the so-called 'first generation' wave of structural reforms (Edwards 1995; Lora 2001). By 1995, Latin American countries had lowered their tariffs from an average of 50 % in the mid-1980s to 12 % in the mid-1990s, and dismantled quantitative restrictions and other forms of non-tariff protection (Franko 2007). Contrary to the advice of those who supported a gradual and sequential approach, particularly in relation to the liberalisation of the capital account, the political economy of the process often led to the bundling of the reforms. But countries pursued the reform process at different speeds. 'Aggressive reformers' (Argentina, Bolivia, Peru, and Chile in an earlier period) often opted for shock therapies, while 'cautious reformers' (Brazil, Colombia, Costa Rica and Mexico) proceeded much more gradually (Stallings and Peres 2000).

To mobilise political support, the benefits of the reform process were oversold (Kingstone 2011). Outcomes in terms of economic growth and social progress were generally disappointing: Although Latin America's annual per capita GDP growth of 1.5 % during the 1990s was much better than the negative growth experienced in the 1980s, it was too low to reduce poverty in a significant way. By the end of the 1990s the rate of poverty still had not been reduced to its 1980 level (ECLAC 2006).

The demise of the Washington Consensus as a development model resulted from the crisis that hit the region in the aftermath of the 1997 Asian financial crisis. The reforms of the 1990s had made Latin America more vulnerable to external shocks. Rather than developing a framework to deal with greater exposure to risk, the region continued to opt for policies that resulted in low domestic savings, high levels of external debt (in sectors with little capacity to generate foreign exchange), overly rigid exchange rates and imprudent lending practices. A sudden reversal in the direction of capital flows following the Asian financial crisis forced an adjustment. The twin fiscal and current account deficits became unsustainable, and as a result currencies were depreciated or allowed to float, and private and public expenditures reduced. Average annual growth per capita was zero between 1998 and 2002 in the region, and unemployment rates rose in most countries, exceeding 20 % in the case of Colombia. The Washington Consensus was blamed and became a politically 'damaged brand'.

In spite of the demonisation of the Consensus, many countries in the region responded to the crisis by strengthening fiscal discipline, which was a core item on Williamson's original list. A number of countries passed 'fiscal responsibility laws' and improved their budgetary institutions. Many introduced new sources of revenue, with innovative (although inefficient) taxes such as those on financial transactions. Some countries, notably Chile, adopted 'fiscal rules' with targets in terms of structural budgets which exclude the transitory components of revenues and expenditures. This allows fiscal policy to operate in a countercyclical manner, running deficits whenever the economy performs below its medium-term trajectory and surpluses when it is above. This feature, together with the adoption of flexible exchange rates in most countries, turned out to be crucial to offset the negative impacts of the global recession of 2008–2009. The efforts that the region undertook to strengthen prudential regulation and de-dollarise financial markets after the banking crises of the late 1990s were similarly critical for navigating the subsequent crisis. (The chapters included in Lora (2007) provide details on each of the reforms adopted since the 1980s.)

Another item that received increased attention after the 1998–2002 crisis was the re-prioritisation of social expenditures, not only to accelerate progress in these areas, but also to offset the consequences of external shocks. Countries throughout the region introduced conditional cash transfer programs (CCTs), such as *Bolsa Familia* in Brazil and *Progresa/Oportunidades* in Mexico, targeted to low-income beneficiaries. These and other innovative programs have been effective at reducing current poverty, and have sought to break the intergenerational transmission of poverty by increasing levels of education, nutrition, and health among the young (Fiszbein and Schady 2009).

Despite these adjustments, the Washington Consensus is no longer regarded as a development model (Birdsall et al. 2010). At best, it is seen as a necessary but insufficient condition for development that highlights the importance of macroeconomic stability. At worst, it is seen as an obstacle to be removed. This is the case in some countries that have reversed trends in market deregulation, liberalisation of foreign direct investment and privatisation. For example, nationalisations of 'strategic' sectors have been a central element of the development agenda in Bolivia, Ecuador and Venezuela. Most countries in the region, though, remain relatively open to trade and capital flows. The most significant change relative to the original Washington Consensus is that the state is back on the development agenda through an active use of what are now called 'productive development policies' (Melo and Rodríguez-Clare 2007).

The emergence of China as a major trading partner – and competitor – is perhaps the most significant economic development in Latin America since 2000. It also helps to explain the re-emergence of productive development policies in the region. In 2010, China accounted for close to one quarter of total exports for Chile, 15 % for Peru and 13 % for Brazil. China's exports to the region have also increased markedly (20-fold in the case of Brazil between 2000 and 2010). Exports from Latin American countries to China are heavily concentrated in primary products, while imports from China include a diverse set of goods, ranging from textiles to sophisticated manufactured products. While some countries have been left out of the expansion of exports to China, almost all have experienced the effects of greater manufacturing imports from China and greater competition in third-party markets, with a cost in terms of output and employment. Indeed, the re-emergence of productive development policies in many countries owes a great deal to the need to curb the process of deindustrialisation and the loss of export market share in products other than commodities.

Institutions and State Capacity

The broad topic of institutional and state reforms, labelled as 'second generation' reforms, proceeded on a separate track and often predated the Washington Consensus (Lora 2007). This is the case, for example, of fiscal and administrative decentralisation and judicial reform that were triggered by the return to democracy and the wave of constitutional reforms which began during the 1980s.

The region has moved forward on a variety of fronts, such as the independence of the judiciary, the professionalisation of the bureaucracy and the quality of political institutions, although the personalisation of politics and the lack of institutionalised and programmatic political parties remains a critical weakness. Specific measures of the quality of the state, however, such as the ability to protect investors against expropriation risk and the World Bank's governance and ease of doing business indicators, still lag behind the developed world and emerging Asia.

Few measures of state capacity are as relevant as the ability of governments to collect tax revenues. In this case, international comparisons are not favorable to Latin America either. Using IMF data for the 1980–2006 period, total tax revenues relative to GDP were 13.4 % in Latin America in contrast to 22.2 % for the world. When only other former colonies are used as a reference, Latin America falls 5.5 points behind. The comparisons are even more telling with income taxes. In this case, the region is percentage points behind the world average of 9.8 % of GDP (Cárdenas 2010).

Interestingly, this is still true despite an active tax reform agenda in Latin America. New taxes have been introduced, but their effects have only offset the decline in revenues associated with lower tariffs and the reduction in corporate income tax rates forced by globalisation.

Of course, generalisations do not capture the varied experiences within the region. A few countries – including Chile and Brazil – have been effective in raising taxes. Consequently, higher fiscal capacity has allowed these countries to provide more public goods and to pursue developmental goals more aggressively. In contrast, the majority of countries in Central America and a number of them in South America, such as Paraguay, have very weak state capacity with tax revenues as a share of GDP in the single digits. As a result, these countries are particularly vulnerable to economic and other shocks, such as natural disasters or security challenges such as the ones confronted in recent years by a number of Central American countries. Weak fiscal capacity also reduces the ability of countries to break out of poverty traps.

Inequality, both economic and political, has been singled out as a crucial obstacle to investment in state capacity (Sokoloff and Zolt 2006). Groups in power prefer the status quo of low taxation, low provision of public goods and low redistribution, perpetuating the effects of extractive colonial institutions. As argued by several authors in the volume edited by Fukuyama (2008), breaking that cycle is one of Latin America's biggest challenges. Progress in terms of democratic institutions is undoubtedly a sign of hope. However, the evidence suggests that the adoption of the formal architecture of democracy does not necessarily deliver the expected results in terms of building state capacity, in part because it is a slow process, but also because economic inequality prevents democratic governance from delivering its full potential.

The Persistence of Poverty and Inequality

High and persistent inequality is perhaps the most salient feature of Latin America's development process. The distributions of income, land, education, health and access to basic services all show extremely high degrees of concentration. Inequality is at the centre of virtually all explanations concerning the region's development problems: economic and social exclusion, limited intergenerational mobility, and weak institutions. However, specific channels and historical evidence are still a matter of controversy.

Although there is some debate about the degree of inequality in Latin America before 1900 relative to other pre-industrial societies or to industrialising Europe, inequality at the time of independence was much higher in Latin America than in North America, mainly as a result of the patterns of colonial land occupation (Engerman and Sokoloff 1997). Recent research suggests that the level of inequality in the region increased after 1870 as a result of the increase in land and mineral rents (relative to wages) during the first phase of commodity export-led growth (Williamson 2010). High inequality in Latin America persisted during much of the twentieth century, in contrast to the equalising trends observed in the industrialised world, explaining why the region has had relatively high levels of inequality compared to Western Europe and Asia (Deininger and Squire 1998). Latin America missed the 'egalitarian revolution' that characterised Western and Eastern Europe, as well as North America and Australia up until the 1980s. This is what distinguishes Latin America from the rest of the world. The fact that Latin America adopted similar labour market and social security institutions suggests that a more fundamental element, namely state capacity, was missing.

A systematic analysis of inequality trends in Latin America is limited by the availability of comparable household surveys. Surveys included in the 2008 World Income Inequality database of UNU-WIDER which cover the time period 1867–2006 differ along many dimensions (coverage area, surveyed population, unit of analysis and measure of welfare). Despite data limitations, the general view is that inequality in Latin America increased during the first half of the twentieth century, and then fell during the 1960s and 1970s, only to increase again during the crisis

of the 1980s. By the mid-1990s, average inequality in the region was comparable to what it had been in the early 1970s (Londoño and Székely 2000). The more recent evidence suggests that inequality increased somewhat in the 1990s as a result of the reforms and then the crisis at the decade's end, but then fell during the 2000s (Gasparini and Lustig 2011, and the country papers in López-Calva and Lustig 2010). This suggests that during the last 40 years inequality trends have often moved inversely to overall economic conditions, and that Latin America made no sustained progress in reducing income inequality. It also suggests that some interventions, such as the slow but important increases in educational achievements in recent decades, take time to produce social dividends. A number of authors have pointed to the expansion of human capital and the associated reductions in wage premia as one of the important reasons for falling inequality in the 2000s (Barros et al. 2010; Székely and Sámano 2011). The rise of conditional cash transfer programs in many countries since the late 1990s also contributed to this decline.

However, there are significant differences across countries. The Gini coefficient for the distribution of national household income per capita ranges between 0.45 in Uruguay and 0.60 in Bolivia. In the case of urban areas and narrower definitions of household income the range goes from 0.45 in El Salvador to 0.55 in Brazil, which is still substantial. Regardless of the measure, Uruguay, Venezuela, Argentina and Costa Rica have relatively low levels of inequality, while Bolivia, Brazil and Colombia are among the most unequal societies in the region. In terms of changes, inequality has fallen significantly in Brazil and Chile since the early 2000s. Mexico has made continued, albeit slow, progress in the reduction of inequality since the early 1990s.

From an analytical perspective, Latin America is characterised by 'excess inequality', meaning that the level of inequality is greater than what would be expected given the level of overall development. The Gini coefficient is around 10 points higher for the average country in the region relative to what would be predicted by a regression on per capita GDP. In terms of fundamental explanations, the dependence on primary activities, the institutions associated with this economic structure, as well as race and ethnic inequalities, are all interdependent forces difficult to isolate (De Ferranti et al. 2004). By any standard, indigenous and afro-descendent groups in Latin America, representing in some countries large shares of the population, are at a disadvantage relative to whites. In contrast, gender gaps have generally narrowed, and in many countries women now obtain more schooling than men.

High levels of inequality contribute to high rates of poverty in Latin America. Some authors estimate that poverty would fall by half if there were no excess inequality in the region (Londoño and Székely 2000). Poverty fell during the period of growth and falling inequality in the 1970s, but then rose sharply during the crisis of the 1980s. Poverty reduction was disappointing in the 1990s, but once again accelerated as income grew and inequality fell in the 2000s. The expansion of conditional cash transfer programs since the late 1990s contributed to this outcome. Poverty fell from 44 to 33 % of the population in the region between 1999 and 2008, and extreme poverty fell from 19 to 13 % (ECLAC 2010).

Until the debt crisis of the 1980s, most of the poor in Latin America lived in rural areas. Since around 1990 poverty became predominantly an urban phenomenon. Poverty rates in rural areas, however, remain almost twice as high as in urban areas – 52 % vs. 27 % in 2008 – and the depth of poverty is more severe. Extreme poverty is three to four times as prevalent in rural areas, which implies that more of the extreme poor in Latin America live in rural areas. As the millennium development goals have gained importance, the first goal of halving extreme poverty and hunger has placed renewed policy attention on rural poverty. Renewed growth and the expansion of CCTs helped push extreme poverty in rural areas down from 38 % to under 30 % between 1999 and 2008.

The causes of rural poverty relate to inequality, institutional deficiencies and market failures. Efforts to reverse high degrees of concentration in land ownership have failed for the most part. Even in Brazil, where there has been an active

land reform program in recent decades, the land Gini remained constant at 0.85 between 1985 and 2006 (Hoffmann and Ney 2010). Insecure property rights make Latin America one of the regions in the world with the lowest share of land rented or sharecropped, which impedes access to land for the landless (de Janvry et al. 2001). Credit market failures create obstacles to buying land, and make it difficult for small farmers to purchase capital and technology that could enhance their productivity. Insufficient land and capital, combined with low levels of education, help explain the high levels of extreme poverty in rural Latin America. For a significant share of the rural poor, poverty reduction will require access to higher productivity wage labour, migration or anti-poverty programs such as CCTs.

Low levels of education are another important explanation for poverty in Latin America. While primary education practically has been universalised, and over 70 % of youth are enrolled in secondary school in many countries, the distribution of education continues to be very unequal. In a number of the poorer countries, such as Nicaragua and Guatemala, net enrolment rates in secondary school are still below 50 %. And in countries like Brazil and Mexico that have much better average outcomes, the distribution is problematic. In Brazil in the mid-1990s, for example, adults in the bottom 40 % of the income distribution had less than half the education of adults in the top 20 % (Székely and Montes 2006). Mexico was little different.

In terms of Latin America's ability to compete in higher value-added activities with East Asian and other developing countries, not only has the rate of improvement in educational attainment lagged, but the quality of education is insufficient. Learning outcomes are captured in the OECD Program for International Student Assessment (PISA) test scores. All Latin American countries that took the PISA exams in 2009 obtained scores that were statistically significantly below the OECD mean, and in some cases were among the four lowest performers in the group of 31 non-OECD countries/economies that participated. Equally problematic is that there are significant quality differences within countries. There is a close association between differences in the socioeconomic background of secondary school students of private *vis à vis* public institutions and the differences in average PISA test scores. The differences in test scores and socioeconomic background of students in Latin America are much greater than those of other developing nations as well as OECD countries. Students in the private system on average perform better than those in the public system. A student in the private system in Brazil, for example, has cognitive skills that are approximately comparable to almost three additional years of public education (OECD 2010). This is a very powerful force that reproduces inequality.

Conclusions

Latin American countries have made enormous strides since the 1870s in improving living standards and human development indicators. Income per capita has increased by a factor of ten, life expectancy has risen by 45 years since 1900, and illiteracy has been reduced from well over two thirds of the population to under 10 %. Latin America has been transformed from a largely rural, agricultural region into a place where 80 % of the population lives in urban areas, over 90 % has access to improved drinking water and 70 % of adolescents attend secondary school.

But the long view on absolute improvements in living standards hides several very different periods. From 1870 to 1981 – a period spanning outward and inward oriented development strategies – Latin America was among the fastest growing regions in the world. Income per capita rose from around 80 to 120 % of the world average, grew as quickly as income in the USA, and rose from about 125 to 280 % of average income in Asia. Yet by 2008, income ratios relative to Europe and Asia were almost identical to what they had been in 1870, and had retreated from 28 down to 22 % of the US level. Latin America has had a growth problem since the early 1980s.

The roots of Latin America's growth problem began well before the debt crisis of the 1980s. In 1960, output per worker was more than one and a

half times greater in Latin America than in East Asia; it is now 50 % smaller. Total factor productivity (TFP) for Latin America relative to the USA, the G8 and East Asia has been declining since the 1960s in most countries in the region. Low TFP can be the result of many forces. One factor that has been singled out is the structure and composition of output in Latin America, which is still characterised by dependence on primary commodities and relatively small domestic markets as a result of the high levels of poverty and inequality. But low productivity also reflects the high levels of informality, as well as low levels of expenditure on innovation, and research and development activities.

Long-term growth that can generate sustained improvements in living standards in Latin America will require gains on many fronts: more effective policies, improved state and market institutions, and a more educated labour force. Another fundamental issue is Latin America's low saving rates. Fortunately, many countries in the region have already begun to make progress on these issues. A new awareness has spread throughout the region about the importance of fiscal responsibility, low inflation and external balance. Simultaneously, most counties have adopted flexible exchange rates, and foreign exchange reserves rose to unprecedented levels in the 2000s. These factors contribute to reducing the impact of external shocks, and were a key reason why the global recession of 2008–2009 was much less painful for Latin America than the impact of the Asian Crisis in the late 1990s.

Latin American countries missed the opportunity that a number of East Asian countries seized in the 1960s and 1970s to invest heavily in education, open their economies and shift into higher value-added manufactured exports. A new group of Asian countries, led by China, has embarked on this path more recently, making this a more contested strategy to pursue. The growth of manufacturing in Asia has created new opportunities by increasing the demand for commodities and food, but it is also prematurely de-industrialising Latin America. The period since 2002 was one of rapid growth in many Latin American countries, and this created a sense of optimism in the region. But it is still too soon to proclaim that a new period of sustained growth has begun.

Latin American countries have always had an abundance of natural resources, and they are well positioned to take advantage of the most recent commodity boom. But they need to do so wisely, lest they repeat past episodes of disequalising growth, with booms followed by painful busts. They also need to find ways to move up the value-added ladder and develop institutions that assist small and medium enterprises to participate in these markets in order to democratise the benefits of export growth.

There are grounds for optimism about the future of the region, but many challenges remain. Latin American countries have accelerated progress on poverty reduction in the past decade as a result of more rapid growth, falling inequality and social policy innovations that hold the promise of reducing the intergenerational transmission of poverty. Primary school has almost been universalised, and enrolments in secondary school have expanded rapidly. Yet the quality of Latin American schools remains extremely low. Educational attainment as measured by years of schooling is important, but if Latin American workers are going to compete successfully, governments in the region must simultaneously prioritise an improvement in learning outcomes.

See Also

- ▶ Development Economics
- ▶ Growth and Institutions
- ▶ Poverty Alleviation Programmes
- ▶ Structuralism
- ▶ Washington Consensus

Bibliography

Baer, W. 1972. Import substitution industrialization in Latin America: Experiences and interpretation. *Latin American Research Review* 7(2).

Barros, R. P. de, de M. Carvalho, S. Franco and S. Mendonça. 2010. *Determinantes da Queda na Desigualdade no Brasil*. IPEA Texto para Discussão 1460, January. Rio de Janeiro.

Birdsall, N., de la A. Torre and F.V. Caicedo. 2010. The Washington consensus: Assessing a damage brand. *World Bank policy research working paper 5316*.

Bulmer-Thomas, V. 2003. *The economic history of Latin America since independence*, 2nd ed. Cambridge: Cambridge University Press.

Cárdenas, M. 2010. State capacity in Latin America. *Economia* 10(2): 1–45.

Cardoso, F.H., and E. Faletto. 1969. *Dependencia y Desarrollo en América Latina; Ensayo de Interpretación Sociológica*. Mexico: Siglo Veintiuno Editores.

De Ferranti, D., G.E. Perry, F.H.G. Ferreira, and M. Walton. 2004. *Inequality in Latin America: Breaking with history*. Washington, DC: World Bank.

de Janvry, A., G. Gordillo, J.-P. Platteau, and E. Sadoulet. 2001. *Access to land, rural poverty, and public action*. New York: Oxford University Press.

Deininger, K., and L. Squire. 1998. New ways of looking at old issues: Inequality and growth. *Journal of Development Economics* 57: 159–187.

Diaz-Alejandro, C.F. 1984. Latin America in the 1930s. In *Latin America in the 1930s: The role of the periphery in world crisis*, ed. R. Thorp. New York: St Martin's Press.

Dos Santos, T. 1970. The structure of dependence. *American Economic Review* 60(2): 231–236.

ECLAC. 2004. *Statistical yearbook for Latin America and the Caribbean, 2003*. Santiago: United Nations.

ECLAC. 2006. *Social panorama of Latin America 2005*. Santiago: United Nations.

ECLAC. 2010. *Statistical yearbook for Latin America and the Caribbean, 2009*. Santiago: United Nations.

Edwards, S. 1995. *Crisis and reform in Latin America: From despair to hope*. New York: Oxford University Press.

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 histories of Brazil and Mexico, 1800–1914*, ed. S. Haber, 260–304. Stanford: Stanford University Press.

Evans, P. 1979. *Dependent development: The alliance of multinational, state, and local capital in Brazil*. Princeton: Princeton University Press.

Ffrench-Davis, R. 2002. *Economic reforms in Chile: From dictatorship to democracy*. Ann Arbor: University of Michigan Press.

Fishlow, A. 1986. Latin American adjustment to the oil shocks of 1973 and 1979. In *Latin American political economy: Financial crisis and political change*, ed. J. Hartlyn and S.A. Morley, 54–84. Boulder: Westview Press.

Fiszbein, A. and N. Schady. with F. Ferreira, M. Grosh, N. Keleher, P. Olinto and E. Skoufias. 2009. *Conditional cash transfers: Reducing present and future poverty*. Washington, DC: World Bank.

Franko, P. 2007. *The puzzle of Latin American development*, 3rd ed. Lanham: Rowman and Littlefield Publishers.

Fukuyama, F. (ed.). 2008. *Falling behind: Explaining the development gap between Latin America and the United States*. New York: Oxford University Press.

Gasparini, L. and N. Lustig. 2011. *The rise and fall of income inequality in Latin America*. CEDLAS Documento de Trabajo No. 118, May.

Hirschman, A.O. 1968. The political economy of import-substituting industrialization in Latin America. *Quarterly Journal of Economics* 82(1): 1–32.

Hirschman, A.O. 1987. The political economy of Latin American development: Seven exercises in retrospection. *Latin American Research Review* 22(3): 7–36.

Hoffmann, R., and M.G. Ney. 2010. Evolução recente da estrutura fundiária e propriedade rural no Brasil. In *A Agricultura Brasileira: Desempenho, Desafios, e Perpsectivas*, ed. J.G. Gasques, J.E.R.V. Filho, and Z. Navarro. Brasília: IPEA.

Jenkins, R. 1984. *Transnational corporations and industrial transformation in Latin America*. New York: St. Martin's Press.

Kay, C. 1989. *Latin American theories of development and underdevelopment*. London/New York: Routledge.

Kingstone, P. 2011. *The political economy of Latin America: Reflections on neoliberalism and development*. New York: Routledge.

Londoño, J.L., and M. Székely. 2000. Persistent poverty and excess inequality: Latin America, 1970–1995. *Journal of Applied Economics* III(1): 93–134.

López-Calva, L.F., and N. Lustig (eds.). 2010. *Declining inequality in Latin America: A decade of progress?* Washington, DC: Brookings Institution Press/UNDP.

Lora, E. 2001. *Las reformas estructurales en América Latina: qué se ha reformado y cómo medirlo*, IADB research department working paper 462. Washington, DC: Inter-American Development Bank.

Lora, E. (ed.). 2007. *The state of state reform in Latin America*. Washington, DC: Inter-American Development Bank/Stanford University Press.

Madison, A. 2008. *Historical statistics of the world economy: 1–2008 AD*. Data downloaded from http://www.ggdc.net/MADDISON/oriindex.htm, June 2011.

Melo, A., and A. Rodríguez-Clare. 2007. Productive development policies and supporting institutions in Latin America and the Caribbean. In *The state of state reform in Latin America*, ed. E. Lora. Washington, DC: Inter-American Development Bank/Stanford University Press.

Milanovic, B. 2011. *The Haves and the Have-Nots: A brief and idiosyncratic history of global inequality*. New York: Basic Books.

Morley, S.A. 1995. *Poverty and inequality in Latin America: The impact of adjustment and recovery in the 1980s*. Baltimore: Johns Hopkins University Press.

Nugent, J.B., and J.A. Robinson. 2010. Are endowments fate? *Revista de Historia Económica* 28(1): 45–82.

OECD. 2010. *PISA 2009 results: Overcoming social background – Equity in learning opportunities and outcomes (Volume II)*. Paris: OECD.

Prebisch, Rául. 1950. *The economic development of Latin America and its principal problems*. New York: The

United Nations. (First published in Spanish by ECLAC in 1949).
Sachs, J.D. (ed.). 1990. *Developing country debt and economic performance, volume 2. Country studies—Argentina, Bolivia, Brazil, Mexico*. Chicago: University of Chicago Press.
Sokoloff, K.L., and E.M. Zolt. 2006. Inequality and taxation: Evidence from the Americas on how inequality may influence tax institutions. *Tax Law Review* 59: 201–276.
Stallings, B., and W. Peres. 2000. *Growth, employment and equity, the impact of the economics reforms in Latin America and the Caribbean*. Washington, DC: Brookings Institution Press/United Nations Economic Commission for Latin America and the Caribbean.
Székely, M., and A. Montes. 2006. Poverty and inequality. In *The Cambridge economic history of Latin America*. New York: Cambridge University Press.
Székely, M. and C. Sámano. 2011. *Trade and income distribution in Latin America: Is there anything new to say?* Mimeo.
Thorp, R. 1998. *Progress, poverty and exclusion: An economic history of Latin America in the 20th century.* Washington, DC: Inter-American Development Bank.
Williamson, J. 1990. Ch. 2: What Washington means by policy reform. In *Latin American adjustment: How much has happened?* ed. J. Williamson. Washington, DC: Institute for International Economics.
Williamson, J.G. 2010. Five centuries of Latin American income inequality. *Revista de Historia Económica* 28(2): 227–252.

Lauderdale, Eighth Earl of [James Maitland] (1759–1839)

Morton Paglin

Keywords

Böhm-Bawerk, E. von; Consumer choice; Corn Laws, free trade and protectionism; Division of labour; Labour theory of value; Lauderdale, Eighth Earl of; Malthus, T.R.; Monopoly; Natural harmony of interests; Over-investment; Parsimony; Period of production; Profit and profit theory; Ricardo, D.; Saving and investment; Sinking fund; Smith, A.; Total output theory; Value

JEL Classifications

B31

Born into a Scottish aristocratic family, Lauderdale entered the House of Commons at the age of 21 as a supporter of the Liberal Whig leader Charles Fox. Following the death of his father, he entered the House of Lords in 1790, where he became known for his defence of civil liberties. After a visit to France in 1792 he publicly expressed sympathy for the ideals of the French Revolution and supported a motion in Parliament (1795) to make peace with the new government of France. In his middle years he swung over to the Tory side and adamantly opposed most economic and political reform measures, especially bills to protect labour (even one which would restrict the use of young children in cleaning chimney flues). His views covered the political spectrum: in 1792 he flirted with Jacobinism, becoming a founding member of the Friends of the People; 40 years later he worked against the Reform Bill of 1832. He died in 1839 at 80, a ripe age indeed for a man known for his apoplectic temper.

Lauderdale had a sustained interest in trade policy, but here he also shifted ground. In 1804 he argued 'that all impediments thrown in the way of commercial communication, obstruct the increase of wealth' (1804, p. 365). Yet in his pamphlet *A Letter on the Corn Laws* (1814) he claimed Adam Smith was in error, and advocated protection for agriculture, a position which he strongly held in the House of Lords for some 20 years.

Apart from some tracts on currency questions and debt policy, Lauderdale's contributions to economic thought are found in one major work, *An Inquiry into the Nature and Origin of Public Wealth* (1804). A second edition (1819) contained only minor revisions. This suggests that Lauderdale's involvement with economic theory was a one-time affair. The intellectual ferment generated by Ricardo's *Principles* (1st and 2nd editions), and the earlier tracts by Malthus, Edward West, and Ricardo on rent and profits seems to have passed him by: no mention of his contemporaries or the theoretical issues which they raised appeared in his new introduction or in the footnotes to the 1819 edition. The focus of both editions is the *Wealth of Nations,* and a large

part of the *Inquiry* is given over to a negation of Smith's conclusions. Specifically, Lauderdale asserts that: (1) the maximization of private riches does not lead to maximum public wealth and welfare; (2) labour is not the cause of value or an adequate measure of value; (3) division of labour is not a major factor in economic growth; (4) parsimony and saving are frequently a public detriment as they may lead to over-investment and a capital glut; and (5) government tax revenues applied to rapid debt reduction ('a forced conversion of revenue into capital') will reduce aggregate consumption, deflate profits and capital values, and result in economic distress.

In developing these ideas Lauderdale exposes his deficiencies as a thinker. His analysis is sketchy, his style prolix and repetitious, and his conclusions based on weak or incomplete reasoning occasionally seem pretentious. Not surprisingly, his contemporaries focused on these flaws. Henry Brougham wrote a long very critical commentary on Lauderdale's *Inquiry* in the *Edinburgh Review* (July 1804), to which Lauderdale responded with an acerbic but not too effective pamphlet. Ricardo exposed several of his logical errors (Ricardo 1823, pp. 267–77, 37ln., 384–5), and Malthus, who on a number of issues (capital glut, value theory and agricultural protectionism) was his intellectual heir, failed to acknowledge his intellectual debt; instead, he accused Lauderdale of 'going too far' in his condemnation of parsimony and savings (Malthus 1836, p. 314), even though, as we shall see, their arguments were quite similar. Despite the negative opinions of his contemporaries, and his modest theoretical ability, Lauderdale now occupies a firm, albeit secondary, place in the history of economic doctrine. We may ask why.

The answer I believe lies in the fact that Lauderdale had a number of valuable insights into the workings of the economy which later economists thought important. Böhm-Bawerk considered Lauderdale's theory of profit a limited but significant step towards the true and complete explanation of interest and profit (that is, his own theory). Following the appearance of Keynes's *General Theory* there was a re-examination of earlier writers who might have anticipated Keynesian ideas on saving, investment and employment. Malthus obviously was placed in the centre of this pantheon of economists, and Lauderdale as an earlier thinker espousing similar ideas was accorded lesser status. This is not a wholly satisfactory way of evaluating past intellectual contributions, but there is no doubt that each age searches for harmonious resonances in the historical literature. Here I shall try to broaden the perspective.

In the *Inquiry* Lauderdale challenged the natural harmony of interests propounded by Smith; namely, that individuals seeking private riches would lead a nation to maximize public wealth. To destroy this identity, Lauderdale tried to prove that the sum of private riches could increase while public wealth and welfare declined. Unfortunately, Lauderdale obfuscated the problem by treating the individual riches occasionally produced by monopoly or a sudden scarcity of supply as a *net* addition to aggregate riches when it was clear that Adam Smith meant aggregate riches in real terms, so the scarcity-induced gains of some are more than offset by real losses of others. Furthermore, Lauderdale overlooked Smith's postulate of free competition as a necessary condition for the coincidence of private and public interest. Ricardo came to Smith's defence and cleared up Lauderdale's ten pages of confusion in a couple of succinct paragraphs (Ricardo 1823, p. 276).

But something positive came out of Lauderdale's discussion of value and riches. His examination of the effect of monopoly on total revenue led to an early and fairly sophisticated discussion of demand curves. Lauderdale reviews empirical estimates of the relationship between a percentage change in the price of a good and the percentage change in the quantity demanded, and notes that for various kinds of consumer goods elasticities may differ. In addition to the concept of price elasticity, Lauderdale gave us the beginnings of a theory of consumer choice, noting the utility sacrifices involved in giving up alternative bundles of goods when consumers make new choices in response to price changes (1804, pp. 59–86). Not surprisingly, Lauderdale rejected the labour theory of value, both as a cause of value and a measure of value (1804, p. 12). Although he

related consumer preferences to demand, and was aware of demand in the schedule sense, he failed to relate costs to supply, and hence his theory of value suffered the inadequacies of all the early supply and demand theories, a weakness which Ricardo pointed out (1823, pp. 384–5).

We now come to the section of the *Inquiry* which has been of most interest in the post-Keynesian period: that dealing with saving, investment and fiscal policy. Lauderdale argues that the social benefits from savings have distinct limits: 'In every state of society, a certain quantity of capital, proportioned to the existing state of knowledge of mankind, may be usefully and profitably employed.' Invention may enlarge the scope for the application of capital, but outlets for profitable investment are still limited by the demand for consumer goods (Lauderdale 1804, p. 227).

Individual parsimony may be misguided, but the harm it does tends to be offset by the prodigality of others. However, when a belief in parsimony leads to bad legislation such as a mandated sinking fund, which forces an increase in public parsimony through taxation and debt reduction, then the results may be 'fatal to the progress of wealth' (Lauderdale 1804, pp. 228–30, 271). But there remains the question of what is the mechanism by which high savings rates or forced parsimony become 'fatal to the progress of wealth'. Superficially this discussion of the evils of parsimony has a Keynesian air to it, but actually Lauderdale (and Malthus) go on to describe a situation in which savings *are* invested, and it is over-investment relative to restricted consumption (made lower by taxation) which finally produces a collapse in profitability.

It is noteworthy that both writers developed a model in which productive applications of net additions to the capital stock are dependent on increases in consumption. They both also failed to recognize that for long periods a nation can use part of its investment for further investment – a deepening of the capital structure or, in Böhm-Bawerk's terms, a lengthening of the period of production, certainly an attribute of 19th-century capitalism. Whatever their limitations, it seems clear that the macroeconomic contributions of Lauderdale and Malthus are more closely related to the growth models of the Harrod–Domar type than to a short-run Keynesian analysis in which output drops because savings are *not* invested. Nevertheless, there is a tenuous connection with Keynes when we look at their descriptions of the late phase of the over-investment cycle. For Lauderdale over-investment reduces profits and the value of capital, and the resulting low prices 'discourage reproduction'. When we observe such deflation we ' must be cautious not to mistake for the effects of abundance that which in reality may be only the effect of failure of demand' (Lauderdale 1804, pp. 263–4). Malthus wrote in a similar vein when he pointed to owners of floating capital vainly seeking investment outlets in the glutted capital markets of Europe (Malthus 1836, p. 420).

We may conclude that the Lauderdale–Malthus theory of total output was not for the most part in the Keynesian mould, but surely that is no reason to downgrade it. Both men saw defects in the Smith–Say–Ricardo theory of total output and employment, and they recognized that restricted consumption and high rates of saving and investment could lead to a sectoral imbalance–a glut of capital, falling profits and, finally, a drop in the inducement to invest. In the policy arena, Lauderdale used these insights to oppose tax surpluses and debt reduction in a period of recession (Paglin 1961, pp. 98–107; Lauderdale 1829).

See Also

► Malthus, Thomas Robert (1766–1834)

Selected Works

1797. *Thoughts on finance*. 3rd ed. London: G.G. & J. Robinson.
1804. *An inquiry into the nature and origin of public wealth*. Edinburgh. Reprinted, New York: Augustus M. Kelley, 1962.
1814. *A letter on the Corn Laws*. Edinburgh: A. Constable.
1829. *Three letters to the Duke of Wellington*. London: J. Murray.

Bibliography

Malthus, T.R. 1836. *Principles of political economy considered with a view to their practical application*. 2nd ed. London: William Pickering. Reprinted, New York: Augustus Kelley, 1964.

Paglin, M. 1961. *Malthus and Lauderdale: The anti-ricardian tradition*. New York: Augustus Kelley.

Ricardo, D. 1823. *Principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Laughlin, James Laurence (1850–1933)

Milton Friedman

Keywords

Cost-push inflation; Free silver; Laughlin, J. L.; Quantity theory of money; Veblen, T.

JEL Classifications

B31

Scholar, teacher, monetary reformer and university administrator, Laughlin was born in Deerfield, Ohio of middle-class parents of modest means. A scholarship plus outside work, largely tutoring, enabled him to attend Harvard. After completing his undergraduate study in history, he did graduate work under Henry Adams, receiving a Ph.D. for a thesis on 'The Anglo-Saxon Legal Procedure'. His subsequent academic career, however, was entirely in economics.

From 1878 to 1888 he taught at Harvard, from 1888 to 1890 he was successively Secretary and President of the Philadelphia Manufacturers' Mutual Fire Insurance Company, from 1890 to 1892 Professor of Political Economy and Finance at Cornell, and in 1892 was persuaded by President Harper to become Head Professor of Political Economy at the new University of Chicago, the position he held until he retired in 1916. From 1916 until his death in 1933 he continued his scientific writing and public activities.

Laughlin's scholarly work was almost entirely in the field of money and banking. Much of it, notably his *History of Bimetallism in the United States* (1885), consisted of a thorough and extremely careful presentation of historical evidence on the development of money and monetary institutions. But Laughlin also wrote extensively on monetary and banking theory, and on proposals for monetary reform. His work on these topics was marred by a dogmatic and rigid opposition to the quantity theory of money, an opposition that developed out of his public activities opposing the free silver movement. The proponents of free silver used a crude form of the quantity theory to support their position, which sufficed to render the theory anathema to Laughlin.

Laughlin's attack on the quantity theory had much in common with recent cost- push or structural or supply shock theories of inflation, in emphasizing the role of factors affecting specific goods and services rather than general monetary influences. Then, as now, such theories ran against the major stream of monetary analysis as exemplified in Laughlin's time by the work of Irving Fisher. As a result, his writings on theory have had no lasting influence on economic thought.

According to Wesley C. Mitchell, one of his students,

> Professor Laughlin's indubitable success as a teacher puzzled many who did not pass through his classroom. He was not an original thinker of great power.
>
> He did not enrich economics .... He did not even keep abreast of current developments in economic theory .... He had a prim and tidy mind, which he kept in perfect order by admitting nothing that did not harmonize with the furnishings installed in the 1880's .... Yet he held that a teacher's aim should be 'the acquisition of independent power and methods of work, rather than specific beliefs'.
>
> The very limitations I have listed helped Professor Laughlin to accomplish this aim.... [His] honesty of purpose impelled others to be honest, which meant that doubting students had to work out the reasons for their dissent .... Laughlin forced one to face intellectual conflicts in his own mind and find out where he stood in the world of ideas. That, I have long believed, was the secret of his success in helping so many students of such diverse capacities to make the most of their several gifts. (1941, pp. 879–80)

As monetary reformer, Laughlin was a leading opponent of the advocates of free silver. He wrote, lectured, and campaigned extensively in favour of 'hard money'. In his widely circulated free-silver pamphlet, *Coin's Financial School,* William Hope Harvey used Laughlin as a hard-money foil for the fictional Coin's free-silver argument. That episode terminated in a widely reported public debate in Chicago in 1885 between Laughlin and Harvey.

After the defeat of William Jennings Bryan and the free-silver forces in the presidential election of 1896, financial and commercial interests in the country organized the Indianapolis Monetary Commission to develop proposals for reform of the monetary and banking system. One of the 11 members of the commission, Laughlin was also the author of its extensive final report, which served as an important stepping-stone en route to the Aldrich–Vreeland Act of 1908 and the Federal Reserve Act of 1913. In addition, Laughlin served for nearly two years from 1911 to 1913, on leave from the University of Chicago, as full-time chairman of the executive committee of the National Citizens League, an organization formed to mobilize public opinion in favour of banking reform.

Laughlin's close links with the Republican Party prevented him from playing any public role in the final preparation of the Federal Reserve Act under a Democratic administration. However, he exerted considerable influence behind the scenes through extensive private correspondence with his former student and assistant, H. Parker Willis, who, as banking expert for the House Banking and Currency Committee, has been regarded as primarily responsible for drafting the Act.

Laughlin's most important and lasting contribution was as head of the Department of Political Economy of the new University of Chicago. Though himself a hard-money man of rigidly conservative views, he demonstrated an extraordinary degree of tolerance for divergent views in staffing and guiding the department. At the very outset, he brought with him from Cornell Thorstein Veblen, who remained in the department for 14 years, the longest period Veblen spent at any single university during his stormy career. Veblen served as managing editor of the *Journal of Political Economy,* which Laughlin founded as one of his first acts at Chicago. Laughlin himself was the editor. As John U. Nef wrote in his obituary notice of Laughlin, 'his wide cultural interests combined with his other qualities to enable him to gather about him a more remarkable group of younger men than was to be found in any other economics department in the country and to help these men in making the most of their own gifts.' Nef notes that a very considerable portion of all the men who have made an important mark in economic thought between 1895 and 1930, beginning with Thorstein Veblen and coming down to Jacob Viner (Laughlin's last appointment) were connected at one time or another, as members or students, with the department of political economy. . .. Laughlin frequently chose the best men when they were of very different persuasions from his own. . .. And so it came about that one of the most conservative heads of an economics department in the country had politically the most liberal and economically the least orthodox department. (1934, p. 2)

Laughlin's emphasis on quality rather than ideology was combined with an emphasis on research by his faculty, as well as by graduate students as part of their training. A corollary was his belief in personal teaching as opposed to formal lecturing. These have remained key characteristics of the Chicago Department of Economics from that day to this. In more recent years, as in his day, the department has been widely regarded as a stronghold of proponents of a free-market economy. That reputation was justified in the sense that throughout the period the department had prominent members who held these views and presented them effectively. But they were always a minority. The department has been characterized by heterogeneity of policy views, not homogeneity. The economists at Chicago who held the generally fashionable views – who were 'liberal' in the 20th-century sense – could be matched at other institutions; the ones who were 'liberal' in the 19th-century sense could not be. That, plus the emphasis on economics as a serious scientific subject, capable of being tested by empirical and historical evidence, and of being

used to illuminate important practical issues of conduct and policy, made Chicago economics unique. These were Laughlin's bequest to the department he built.

Selected Works

1885. *The history of bimetallism in the United States*. New York: D. Appleton & Co.
1898. *Report of the Monetary Commission of the Indianapolis convention of boards of trade, chambers of commerce, commercial clubs, and other similar bodies of the United States*. Chicago: University of Chicago Press.
1903. *The principles of money.* New York: Charles Scribner's Sons.
1909. *Latter day problems.* New York: Charles Scribner's Sons.
1931. *A new exposition of money, credit and prices,* 2 vols. Chicago: University of Chicago Press.
1933. *The federal reserve act: Its origin and problems.* New York: Macmillan.

Bibliography

Bornemann, A. 1940. *J. Laurence Laughlin*. Washington, DC: American Council on Public Affairs.
Mitchell, W.C. 1941. J. Laurence Laughlin. *Journal of Political Economy* 49: 875–881.
Nef, J.U. 1934. James Laurence Laughlin (1850–1933). *Journal of Political Economy* 42: 1–5.

Launhardt, Carl Friedrich Wilhelm (1832–1918)

Jürg Niehans

Keywords
Banking School; Exchange; Hotelling, H.; Investment criteria; Jevons, W. S.; Launhardt, C. F. W.; Marginal cost pricing; Marginal revolution; Marginal utility of money; Monopoly; Optimal tariffs; Price discrimination; Quantity theory of money; Rate of interest; Rent; Transportation economics; Wage differentials; Walras, L

JEL Classifications
B31

Launhardt was born on 4 April 1832 in Hannover, where he died on 14 May 1918. His work is Germany's most important and in fact only significant contribution to the 'marginal revolution' in the last three decades of the 19th century. In the economic analysis of transportation and location, this contribution was not surpassed until the 1930s. Available only in German, some of it in publications that are hard to find, it still has not found the recognition it deserves, and Schumpeter's references in the *History of Economic Analysis* are inadequate.

Like Dupuit, Launhardt began his professional life as a civil engineer, working for the public road administration. In 1869 he joined the faculty of the Hannover Polytechnic Institute as a professor for roads, railways and bridges. This was the beginning of a distinguished academic career, in the course of which he served as the director of the institute and, when it became the Technische Hochschule Hannover, its first rector. He was made a member of the Königliche Akademie des Bauwesens and of the Preussische Herrenhaus. Dresden gave him an honorary degree for his contributions to the technology and economics of transportation.

Practical problems of highway planning led Launhardt to the gradually more general analysis of efficient transportation networks. This work was later systematized in *Theorie des Trassirens* (Theory of Network Planning). Part I, entitled 'Commercial Network Planning', contains the derivation of efficiency criteria without regard to topography. This part is the second edition, much revised and enlarged, of the 1872 publication, and also incorporates sections from the 1885 book. Part II, entitled 'Technical Network Planning for Railroads', applies economic efficiency criteria to curves and gradients imposed by topography; an earlier version was published in 1877.

The contributions to economics are found in Part I. This begins with a discussion of investment criteria. From a social point of view, networks should be planned in such a way that the sum of operating and capital costs is a minimum. Private capitalists, however, try to maximize the internal rate of return on their capital. Under perfect competition the two criteria would coincide, since the internal rate of return, if duly maximized, would equal the market rate of interest. In reality, however, since the railroad industry is inherently non-competitive, rates of return can be pushed above market rates of interest by keeping railroad investment below the social optimum. This was one of Launhardt's basic arguments for government ownership of railroads. For his own analysis he uses, of course, the social criterion.

Using geometry and calculus, Launhardt derives rules, depending on freight costs and volumes, for the optimal direction and density of highways connecting given market centres. He shows that highways of different quality (and thus with different freight costs) should meet at angles analogous to those of refracted light, a rule later popularized by Stackelberg as the 'law of refraction'. According to the 'law of nodes', transport costs on a star-shaped transportation network connecting three cities are minimized if the sines of the angles between its rays bear the same proportions as the total transportation costs per mile along the rays. The efficient combination of different modes, like highways, waterways and railways, is also considered.

Applying his analysis of network nodes to the location of plants, Launhardt produced the first substantial theory of industrial location (1882). In this basic contribution he determines the efficient location of a plant with given sources of supplies and given sales outlets by minimizing transportation costs. The optimum is found by an ingenious geometrical construction which became known as the 'pole principle', later amplified by Palander. It is given a mechanical interpretation as the centre of gravity of forces, representing freight rates, acting at the different input and output locations. After first assuming that the network of routes is being planned from scratch, Launhardt also derives rules for optimal additions to existing networks. The analysis is far superior to that in Alfred Weber's later book on the location of industries, in which Launhardt is not mentioned, and whose only claim to attention is the appendix by Georg Pick.

Launhardt's main contribution to the theory of railway rates is found in chapter 32 of (1885). It was elaborated in (1887) and further detail was added in (1890a) and (1890b), but these extensions add nothing for more general economic interest. The paper on 'Economic Problems of the Railway Industry' provides an extensive analysis, based on consumer surplus, of the social rate of return of railroads, both theoretical and numerical, including a cost–benefit analysis of future railway development.

For railway rates, Launhardt establishes the principle that the maximization of social welfare requires – in modern terminology – marginal cost pricing. But this, in turn, requires competition, while profit maximization by monopolistic railway firms implies that rates exceed marginal cost. In particular, if a railway transports homogeneous goods from a uniform plain to a market centre, the monopoly price is calculated to exceed marginal cost by 50 per cent (because, in modern terminology, freight volume reacts to the freight rate with an elasticity of – 2 and ton-miles thus with an elasticity of – 3). As a consequence, the freight volume is suboptimal. By perfect discrimination according to 'what the traffic will bear' over each distance, both railway profits and general welfare can be increased compared with simple monopoly. This, however, is only a second-best solution. For Launhardt, the efficiency of marginal cost pricing is another basic argument for government ownership.

Launhardt's main claim to a prominent place in the history of economic analysis is his slender treatise *Mathematische Begründrung der Volkswirtschaftslehre* (Mathematical Foundations of Economics) of 1885. It was written in the light of Walras's *Mathematische Theorie der Preisbestimmung wirtschaftlicher Güter* (1881) and the second edition of Jevons's *Theory of Political Economy* (1879). At the same time, it is clearly pre-Marshall and pre-Edgeworth (though *Mathematical Psychics* had appeared in 1881). Two other books by Walras, sent by the author, arrived

too late to be of use, nor was Launhardt acquainted with Cournot at that time. He reports that the copy he finally obtained from a library had apparently never been read, and Gossen could nowhere be found (because virtually no copies had been sold). Launhardt shows what a competent engineer with an economic turn of mind and a little calculus could do (and also what he could not do) in economics a hundred years ago. Launhardt's addiction to special functional forms, particularly quadratic utility functions, often results in spurious precision, limited generality and reduced lucidity, but the basic contributions are sound, important and original.

In his theory of exchange (Part I), Launhardt rightly criticizes Walras for believing (if taken literally) that there is no way for a trader to improve his position relative to free competition at uniform prices. His counter-examples relate to monopoly and price discrimination, leading him to the idea of an optimal tariff.

While valid in principle, this analysis falls short of Edgeworth's. The discussion of the total gain from trade and its distribution, whose shortcomings were pointed out by Wicksell, was soon obsolete because of its dependence on the interpersonal additivity of utility.

In his discussion of distributive shares, Launhardt recognizes the backward-bending supply curve of labour and the effect of property incomes on labour supply and thus on wages. He also recognizes that the inter-occupational mobility of labour tends to equalize relative wage rates with both the ratios of the marginal products of labour and (to the extent an individual can choose between occupations) the ratio of its marginal disutilities. For profits, Launhardt's 'basic equation' expresses, substantially, the familiar optimality condition that the profit margin, as a percentage of price, is the inverse of the elasticity of demand (though this concept is not used, of course). It is clearly explained that the entrepreneur, in setting his price, considers only marginal costs, while prices are equalized to the average costs of the marginal firm by exit and entry. The profits of intra-marginal firms are correctly interpreted as rents, and the same principle is used to explain wage differentials.

Launhardt's theory of interest is Jevonian in spirit. Though brief and somewhat sketchy, it anticipates all the basic elements of Fisher's theory. In many respects Launhardt achieves more in 20 pages than Böhm-Bawerk in about 500. Using modern terminology, the rate of interest is explained by the interplay between a psychological preference for present consumption, modified by variations in expected income, and the marginal productivity of capital (ch. 24). Saving is interpreted as a sacrifice of current consumption for the sake of an infinite stream of additions to future consumption. It is shown mathematically that, with a rising rate of interest, given the rate of time preference, saving first rises to a maximum and then declines, because at high interest rates small savings are enough to buy a lot of future income. According to the 'basic principle of accumulation', the present value of the future marginal utility of income is made equal to the current marginal utility of income. In the course of time, optimal saving, if initially positive, will decline until a steady state is reached (ch. 15). Investments will be made up to the point where the marginal saving in operating costs is equal to the rate of interest.

The subject of Part III is the effect of transportation on production and consumption. Launhardt starts out by determining production and prices of a single seller supplying an unlimited market of uniform density. Delivered prices are seen to rise towards the periphery in the shape of a hollow cone, known as the 'Launhardt Funnel' (ch. 27). If sellers of differentiated products compete in a uniformly populated plain, their market areas are shown to be polygons, whose sides, depending on circumstances, are pieces of ellipses, hyperbolas or straight lines. In this context there emerges what Palander later called the Launhardt–Hotelling solution for heterogeneous duopoly. Forty-four years before Hotelling, Launhardt already used the paradigm of two competing suppliers, located at different points along a street, each maximizing his profits on the assumption that the price of his competitor is given. His solution, forgotten for half a century, is substantially identical to Hotelling's. An analogous analysis is provided for suppliers of differentiated

products at the same location, showing how their ring-shaped market areas depend on transportation costs (ch. 29).

From the market areas of given suppliers, Launhardt shifts his attention to the supplying areas of given markets, which brings rent to the foreground. His description of the product 'rings' surrounding a single market city in an unlimited plain adds nothing to Von Thünen (ch. 30). The analysis is then extended to a number of markets, each with its limited supplying area. If identical cities are located in a pattern of regular triangles, the supplying areas are, of course, hexagonal. While this foreshadows Lösch's later work, Launhardt's triangular pattern is based on intuition and not on explicit optimality conditions. It is shown, however, how the mutual limitation of adjoining supplying areas raises rent and product prices (ch. 31). Much of this material was later incorporated in the second edition of *Commercial Network Planning* (1887).

Launhardt's monetary theory is far inferior to his microeconomics. Its centrepiece is the rejection of the quantity theory of money. In part, this is based, in the tradition of Senior and the Banking School, on the argument that under a gold standard an increase in the quantity of paper money just leads to an external (and/or internal) gold drain, while commodity prices remain tied to international prices or, in a closed economy, the gold price. To this extent, Launhardt is on firm ground. He went much further, however. In the theory of relative prices he had assumed that the marginal utility of money is constant. When first introduced, this was an innocuous simplification, but in the theory of money it became the source of fatal confusion, for it induced Launhardt to treat money incomes, which he chose as the proximate determinant of absolute prices, as if they were 'real' variables, independent of the money supply. After that, one is hardly surprised to read that higher interest rates result in higher prices and that gold discoveries have no influence on prices. The basic argument is found in *Mathematische Begründung* (1885); later elaborations (1889; 1894) and historical illustrations and applications.

Selected Works

1868. *Bestimmung der zweckmässigsten Steigungsverhältnisse der Chausseen.* Hannover.
1869. *Ueber Rentabilität and Richtungsfeststellung der Strassen.* Hannover.
1872. *Kommercielle Tracirung der Verkehrswege.* Hannover.
1877. Die Betriebskosten der Eisenbahnen in ihrer Abhängigkeit von den Steigungs-und Krümmungsverhältnissen der Bahn. *Handbuch für specielle Eisenbahn-Technik,* vol. 4, Supplement. Leipzig.
1882. Die Bestimmung des zweckmässigsten Standortes einer gewerblichen Anlage. *Zeitschrift des Vereines deutscher Ingenieure* 26, 106–15.
1883. Wirtschaftliche Fragen des Eisenbahnwesens. *Centralblatt der Bauverwaltung,* vol. 3.
1885. *Mathematische Begründung der Volkswirtschaftslehre.* Leipzig.
1887–8. *Theorie des Trassirens.* 2 parts. Hannover.
1889. *Die Quantitätshtheorie. Ein Beitrag zur Lehre vom Wesen des Geldes.* Hannover.
1890a. *Theorie der Tarifbildung der Eisenbahnen.* Berlin.
1890b. Zur Frage einer besseren Feststellung des Personenfahrgeldes. *Organ für die Fortschritte des Eisenbahnwesens*.
1894. *Mark, Rubel und Rupie Erläuterungen zur Währungsfrage und Erörterungen über das Wesen des Geldes*. Berlin.
1900. Am sausenden Webstuhl der Zeit. *Aus Natur und Geisteswelt* 23. Leipzig.

Laveleye, Emile de (1822–1892)

A. Courtois

Born at Bruges, died at Liège, Laveleye was a remarkable thinker, and his writings were brilliant in style. Unfortunately for his fame, being not only an economist but also a philologist, an

historian, a student of law, a politician, and a moralist, he was scarcely able to fathom the depths of all the subjects he undertook. Absolutely sincere in mind, he allowed himself some inconsistencies of expression which he fully admitted. At one time he frankly acknowledged himself a 'socialist of the chair'; but towards the end of his life the disquieting spectacle of the progress of socialism appeared to draw him nearer to those whom earlier he had stigmatized as 'orthodox economists'.

His principal economic writings are *Le Marché Monétaire et ses crises depuis cinquante ans,* in which he announced himself in favour of the unity and monopoly of banks of issue. *Le socialisme contemporain* (1881); with essay on luxury, etc. (several editions have been published), in which he examined critically the doctrines of Rodbertus, Karl Marx, Ferdinand Lassalle, etc.; and *De la propriété et de ses formes primitives* (1873). He maintained that property was a civil institution, agreeing in this with John Stuart Mill. His last work was *Eléments d'économie politique,* (1882), a text-book on the elements of the science. In monetary questions de Laveleye was a partisan of a double standard, and produced many works supporting bimetallism. He contributed to several periodicals of the day, among others to the *Revue trimestrielle,* to the *Libre recherche,* to the *Revue des deux Mondes,* and to the original edition of this Dictionary the article on Commune. All men of science admired his sincerity, the boldness with which he championed new ideas, his modesty, and his absolute truthfulness. These qualities gave his works an attractive power which won him many readers. The obituary notice of Emile de Laveleye, written by his pupil and successor in the chair of political economy at Liège, Professor Ernest Mahaim, in the *Economic Journal,* Vol. II, speaks of

> the governing idea of his life as being found in the supremacy of justice. He was persuaded that the human race was marching toward an ideal of justice, an image of God, to which ultimately it would attain. He had faith in the boundless progress of mankind, and in the solidarity of all men; and he discerned in the future a society of love, peace, and justice, bringing universal happiness. Emile de Laveleye is a great figure in the century that is passing away.

Professor Mahaim describes de Laveleye as an academic socialist. He believed in the frequent necessity of state intervention to secure the triumph of the common interest over particularist egoism. His criticism, in the *Contemporary Review,* of Mr Herbert Spencer's *The Man* versus *the State,* disclosed how far he repudiated the 'orthodox' *credo.*

> . . . He often sent articles to English newspapers, amongst others to the *Times* and *Pall Mall Gazette.* . . . He had a great affection for England; of its language he had perfect mastery; and on its soil he counted many of the most distinguished politicians among his friends.

Lavington, Frederick (1881–1927)

P. Bridel

L

After eleven years' service in a bank, Lavington went into residence at Cambridge and – together with Dennis Robertson and Hubert Henderson – was among Keynes's first students. After taking his degree in 1911, he returned to administrative work in the then new Labour Exchanges Department of the Board of Trade. Back in Cambridge in 1918, he was elected to a lectureship in economics which he held until his death.

The limited influence Lavington had on the development of Cambridge monetary thought is not difficult to explain. Besides the brevity of his academic career, his belief that 'it's all in Marshall, if you'll only take the trouble to dig it out' (Wright 1927, p. 504) did not induce him to break much new ground. His task – as he saw it – was to apply Marshall's analysis to the practical problems of the money and capital markets.

Seldom cited by his fellow Cambridge economists in his lifetime, Lavington's prescient elaboration of Marshall's cash balance equation was rescued from oblivion by Robertson and Hicks

in their 1937 debate with Keynes on the theory of interest. In his *English Capital Market* (1921, pp. 29–33), Lavington extends the analysis of the demand for money beyond money in the form of income, or transaction deposits and, in particular, takes account for the first time of the influence on this demand of the rate of interest and of the general state of expectations. This piece of analysis clearly anticipates Robertson's 'three-fold-margin' argument and Keynes's liquidity preference doctrine – two of the main stepping stones of the loanable-fund theory of interest ultimately brought to fruition by the latter in his *Treatise on Money* (1930).

Selected Works

Lavington, F. 1921. *The English capital market.* London: Methuen.

Lavington, F. 1922. *The trade cycle.* London: P.S. King & Son.

Bibliography

Wright, H. 1927. Frederick Lavington. *Economic Journal* 37: 503–505.

Law and Economics

David Friedman

The economic analysis of law involves three distinct but related enterprises. The first is the use of economics to predict the effects of legal rules. The second is the use of economics to determine what legal rules are economically efficient, in order to recommend what the legal rules ought to be. The third is the use of economics to predict what the legal rules will be. Of these, the first is primarily an application of price theory, the second of welfare economics, and the third of public choice.

Predicting the Effect of Laws

Of the three enterprises, the least controversial is the first – the use of economic analysis to predict the effect of alternative legal rules. In many cases, the result of doing so is to show that the effect of a rule is radically different from what a non-economist might expect.

Consider the following simple example. A city government passes an ordinance requiring landlords to give tenants three months notice before evicting them, even if the lease agreement provides for a shorter period. At first glance, the main effect is to make tenants better off, since they have greater security of tenure, and to make landlords worse off, since they now find it more difficult to evict undesirable tenants.

The conclusion is obvious; it is also false. The new ordinance raises the demand curve; the price at which tenants choose to rent any given quantity of housing is higher, since they are getting a more attractive good. It also raises the supply curve, since the cost of producing rental housing is now higher. If both the supply and the demand curve rise, so does the price. In the short run, the regulation benefits the tenant at the expense of his landlord. Once rents have had time to adjust, the tenant is better off by the improved security of his apartment but worse off by the higher rent he pays for it; the landlord is worse off by the increased difficulty of eviction and better off by the increased rent he receives.

One can easily construct specific examples in which such a regulation makes both landlords and tenants worse off, by adding to the lease terms which increase the landlord's costs by more than they are worth to the tenant and increase the market rent by more than enough to eliminate the tenants' gain but too little to compensate the landlords' loss. One can also construct examples in which both parties are better off, because the regulation saves them the cost of negotiating terms which are in fact in their mutual interest. Thus economic analysis radically alters the grounds on which the regulation can be defended or attacked, eliminating the obvious justification (helping tenants at the expense of landlords) and

replacing it with a different and much more complicated set of issues.

In this example, and in many similar ones, the two parties are linked by a contract and a price. In such cases, the first and most important contribution of economics to legal analysis is the recognition that a legally imposed change in the terms of the contract will result in a change in the market price. Typically, the result is to eliminate the transfer that would otherwise be implied by the change.

This is not true for cases, such as accidents and crimes, where there is no contract and no price. In analysing such situations, the essential contribution of economics is to include explicitly the element of rational choice involved in producing outcomes that are commonly regarded as either irrational or not chosen.

Consider automobile accidents. While a driver does not choose to have an accident, he does make many choices which affect the probability that an accident will occur. In deciding how fast to drive, how frequently to have his brakes checked, or how much attention to devote to the road and how much to his conversation with the passenger next to him, he is implicitly trading off the cost of an increased risk of accident against the benefit of getting home sooner, saving money, or enjoying a pleasant conversation. The amount of 'safety' the driver chooses to 'buy' will then be determined by the associated cost and benefit functions. Thus, for example, Peltzman (1975) demonstrated that safer autos tend to result in more dangerous driving, with the reduction in death rates per accident being at least partly balanced by more accidents, as drivers choose to drive faster and less carefully in the knowledge that the cost of doing so has been lowered.

This way of looking at accidents is important in analysing both laws designed to prevent accidents, such as speed limits, and liability laws designed to determine who must pay for accidents when they occur. From the economic perspective, the two sorts of laws are alternative tools for the same purpose – controlling the level of accidents.

A driver who knows he will be liable for the costs of any accidents he causes will take that fact into account in deciding how safely he should drive. Elizabeth Landes, in a study of the shift to no-fault auto insurance, concluded that one effect of the reduction in liability was to increase highway death rates by about 10–15 per cent.

The advantage of liability over direct regulation is that the knowledge that if he causes an accident he must pay for it gives the driver an incentive to modify his behaviour in any way that will reduce the chance of an accident, whether or not others can observe it. Regulations such as speed limits control only those elements of driver behaviour which can be easily observed from the outside – speed but not attention, for example. The disadvantage of liability is that it forces drivers, who may well be risk averse, to participate in a lottery – one chance in two thousand, say, of causing an accident and having to pay all of its cost.

An accident is one example of an involuntary interaction; a crime is another. Economic analysis of crime starts with the assumption that becoming a criminal is a rational decision, like the decision to enter any other profession. Changes in the law which alter either the probability that the perpetrator of a crime will be punished for it or the magnitude of the punishment can be expected to affect the attractiveness of the profession, hence the frequency with which crimes occur – as demonstrated empirically in Ehrlich (1972). Similarly, changes in crime rates will, via the rational decisions of potential victims, affect expenditures on defending against crime.

Another area of law, in which the application of economic analysis is less novel, is antitrust. One important contribution of economic analysis has been to suggest that some elements of anti-trust law may be based on an incorrect perception of how firms get and maintain monopoly power.

McGee (1958) used arguments originally proposed by Aaron Director to show that if, as commonly alleged, Standard Oil had attempted to maintain its market position by predatory pricing – cutting the price of oil below cost in order to drive out smaller but equally efficient rivals – the effort would probably have failed. Standard's larger assets would be balanced by a larger volume of sales, and hence larger losses

when those sales were at a price below cost. Even if the smaller firm had gone bankrupt first, its physical plant would have remained, to be purchased by some new competitor. Based on a study of the record of the Standard Oil anti-trust case, McGee concluded that predatory pricing was a myth: Rockefeller had in fact maintained his position by buying out rivals, usually at high prices.

The argument, if correct, implies that some conventional anti-trust activity is misplaced. Pricing policies which are attacked as predatory may in fact be ways in which new firms break into existing markets, using low prices to induce potential customers to try their products. If so, prohibiting such policies reduces competition and encourages the monopoly that the law is intended to prevent.

Efficiency: Prescribing Laws

The use of economic analysis to determine what the law ought to be starts with one simple and controversial premise – that the sole purpose of law should be to promote economic efficiency. There are two problems with this premise. The first is that it depends on the utilitarian assumption – that the only good is human happiness, defined not as what people should want but as what they do want. The second is that economic efficiency provides at best a very approximate measure of what most of us understand by 'total human happiness', since it assumes away the problem of interpersonal utility comparisons by, in effect, treating people as if they all had the same marginal utility of income.

One reply to this criticism is that while few people believe that economic efficiency is all that matters, most people who understand the concept would agree that it is either an important objective or an important means to other objectives. Hence while maximizing economic efficiency may not be the only purpose of laws, it is an important one – and one that economic theory can, in principle, tell us how to achieve. Further, economic theory suggests that an improvement in efficiency may be something that courts can achieve, whereas redistribution, for reasons suggested in the discussion of landlord-tenant relations, may not be.

Once one accepts economic efficiency as the objective, the standard tools of welfare economics can be used to analyse a wide variety of legal issues. Consider, for example, the eviction regulation discussed earlier. If the additional security of tenure is worth more to the tenant than it costs the landlord to produce, then landlords will find it in their interest to include that condition in the lease contract whether or not the law requires them to; the additional rent they will be able to charge will more than make up for the cost of delays in evicting undesirable tenants. If, on the other hand, security of tenure costs the landlords more than it is worth to the tenants, then they will not choose to offer it – and, viewed from the standpoint of economic efficiency, a regulation compelling them to do so is undesirable.

So one conclusion suggested by such analysis is a strong case for freedom of contract – allowing the parties to a lease, or any other contract, to include any terms mutually agreeable. To the extent that one accepts that argument, the function of legal rules is simply to specify a default contract – a set of terms that apply unless the parties specify otherwise. If the default contract closely approximates what the parties would agree to if they did specify all the details of their agreement, it serves the useful purpose of reducing the cost of negotiating contracts.

An important example of such analysis occurs in the case of product liability law. Just as with lease contracts, the first step is to observe that changes in who bears the liability for product defects will produce corresponding changes in market price, so that shifting liability from, say, buyer to seller will not in general result in the buyer being better off and the seller worse off. Changes in liability law will, however, change the incentives facing both buyer and seller with regard to decisions they make that affect the damage produced by defects. To the extent that a buyer cannot judge the quality of a product before he buys it, a rule of *caveat emptor* gives the seller an inefficiently weak incentive to prevent defects, since he pays the cost of quality control and receives no corresponding benefit. On the other

hand, a rule of *caveat venditor* provides the seller with the appropriate incentive, since he ends up paying, via damage suits, for the cost of defects, but it gives the buyer an inefficiently low incentive to try to use the product in way that will minimize the damage from defects – by, for example, driving an automobile in a way that does not rely too heavily on the brakes always working perfectly.

This suggests that different legal rules may be appropriate for different sorts of goods. It also suggests that some intermediate rule, such as contributory negligence, in which the producer of a defective good may defend himself against a damage suit by showing the accident was in part the result of imprudent use by the purchaser, may be superior to both *caveat emptor* and *caveat venditor*.

Just as in the case of tenant and landlord, the analysis suggests that while the law may set a default rule, it ought to permit freedom of contract. Sellers can then convert *caveat emptor* into *caveat venditor* by offering a guarantee, and buyers can convert *caveat venditor* into *caveat emptor* by signing a waiver.

Another area of interest is corporate law. Here the central problem is that of structuring the contract which defines the corporation so as to control the principal-agent problem resulting from the separation of ownership and management. One solution, missed in Smith's classic statement of the problem (Smith 1776), is the takeover bid, used to discipline managers who do not maximize the value of the assets they manage. The question of whether the law should assist or oppose managers in their attempt to prevent takeovers has been a lively issue in the recent literature.

Freedom of contract is of no use where there is no voluntary agreement among the parties. The law must somehow specify who is responsible under what conditions for the cost of accidents, and what the punishment is to be for crimes. One traditional approach to this problem is the 'Hand formula', according to which someone is judged negligent, hence legally responsible for an accident, only if he could have prevented it by precautions that would have cost less than the expected cost (probability times damage) of the accident. This seems to fit very neatly into the economic analysis of law, since it punishes someone only if he has acted inefficiently by failing to take a cost-justified precaution.

It has, however, two serious difficulties. One is that 'accidents' are usually the result of the joint action of two or more parties. My bad brakes would not have injured you if you had not chosen to ride a bicycle at night wearing dark clothing – but your bicycle riding would not have put you in the hospital if my car had had good brakes. In such a situation, the efficient solution is to have precautions taken by whichever party can take them most cheaply – even if the other party could prevent the accident at a cost lower than the resulting damage. This suggests that the Hand formula should be interpreted as making the party liable who could have avoided the accident at the lower cost. Situations in which the probability and cost of accidents are continuous functions of both my level of precaution and yours require additional elaborations of the formula.

A second problem is that the Hand formula requires the court to make judgements, both about the probability of accidents given various levels of precaution and about the cost of both precautions and accidents to the parties involved, which it may not be competent to make. This suggests the desirability of legal rules which are sufficiently general so that they do not depend on a court making case-by-case evaluations of cost and benefit, but which give the parties incentives to use their private knowledge of costs and benefits to produce efficient outcomes. The attempt to construct such rules, for a wide variety of legal problems, makes up a considerable part of the law and economics literature.

Crimes, like accidents, involve involuntary interactions. The economic analysis of crime focuses on two related issues – the incentives facing the criminal and the incentives facing the system of courts and police. The first leads to the question of what combination of punishment and probability of apprehension would be applied, for any crime, in an efficient system; the answer involves trading off costs and benefits to criminals, victims, and the enforcement system. The

second leads to questions about the procedures used by the court system to determine guilt or innocence (also an issue in other parts of the law), and of the relative advantages of private enforcement of law, as in our civil system, in comparison to public enforcement, as in our criminal system.

Economists Learning from Law: The Coase Theorem

So far, all of the example of economic analysis of law have involved using existing economic theory to analyse the law. There is at least one area, however, where the interaction of law and economics has resulted in a substantial body of new economic theory. This is the set of ideas originating in the work of Ronald Coase and commonly referred to as the Coase Theorem.

According to the traditional analysis of externalities associated with Pigou, an externality exists where one party's actions impose costs on another, for which the first need not compensate him. This leads to an inefficient outcome, since the first party ignores the costs to the second in making his decision. Thus, for example, a railroad company may permit its locomotives to throw sparks, even though they cause occasional fires in the neighbouring corn fields. The cost of modifying the engine to prevent sparks would be borne by the company; the cost of the fires is an externality imposed on the adjacent farmers. The traditional solution is a Pigouvian tax. The railroad company is charged for the damage done, and can either pay or stop doing the damage, whichever costs less.

Coase pointed out that in this and many other cases, the cost is not simply imposed by one party on the other, rather, it arises from incompatible activities by two parties. The fires are the result both of the railroad company using a spark-throwing locomotive and of the farmers choosing to grow inflammable crops near the rail line. The efficient solution might be to modify the locomotive, but it also might be to grow different crops. In the latter case, a Pigouvian tax on the railroad leads to an inefficient outcome.

Hence the first step in Coase's analysis suggests that there is no general solution to the problem of externalities. The legislature, in setting up general laws, cannot know which party, in any specific case, will be able to avoid the problem at the lowest cost. If it attempts to solve that problem by a law making whichever party can avoid the problem at the lower cost liable, the court is left with the problem of estimating the costs. Each party has an incentive to misrepresent the cost of its potential precautions, in order to make the other party liable for preventing the damage.

The second step is to observe that both this argument and the traditional analysis of externalities ignore the possibility of agreements between the parties. If the law makes the railroad liable for the damage when the farmers can prevent it at a lower cost, it will be in the interest of both farmers and railroad to negotiate an agreement in which the railroad pays the farmers to grow clover rather than corn along the rail line. Hence this line of analysis leads to the conclusion that whatever the initial definition of rights – whether the railroad has the right to throw sparks or the farmers to enjoin the railroad or collect damages – market transactions among the participants will lead to an efficient outcome.

The final step in the argument is to observe that inefficient outcomes do in fact occur, and that the reason is transaction costs. If, for example, any farmer can enjoin the railroad from throwing sparks, then the railroad, in dealing with the farmers, is faced by a hold-out problem. A single farmer may try to collect a large fraction of what the railroad saves by not modifying its locomotive, using the threat that if his demands are not met he can enjoin the railroad, whatever the other farmers do. If, on the other hand, the railroad is free to throw sparks and it is up to the farmers to offer to pay for the modifications, then in raising the money to do so they face a public good problem; a farmer who does not contribute still benefits. Transaction cost problems of this sort may prevent the process by which bargaining among participants would otherwise lead to an efficient outcome.

The conclusion of all of this is the Coase Theorem, which states that in a world of zero

transaction costs any initial definition of rights will lead to an efficient outcome. It is important not because we live in such a world, but because it shows us a different way of looking at a large range of problems – as resulting from the transaction costs that prevent the parties affected from bargaining their way to an efficient outcome.

This approach represents both an important change in the traditional economic analysis of externalities and a powerful tool for analysing legal institutions. Many such issues can be seen as questions of how property rights are to be bundled. When I acquire a piece of land, does what I buy include the right to make loud noises on it? To prevent passing locomotives from throwing sparks on it? To leave objects lying about that might be hazardous to neighbours who accidentally trespass? From the perspective of the Coase Theorem, all such questions can be approached by asking first what bundling of rights would lead, under various circumstances, to an efficient outcome, and second, if a particular initial bundling of rights leads to inefficient outcomes, how easy will it be for the parties to negotiate a change, with the party who has a greater value for one of the rights in a bundle purchasing it from its initial owner.

One example is the law of attractive nuisance. Does the ownership of a piece of land include the right to put on it open cement tanks full of deadly chemicals, protected only by large signs – which are no barrier at all to a trespasser too young to read? The immediate answer is that the right to decide whether the tanks are fenced is worth more to the neighbourhood parents than to the owner of the property. The further answer is that if the law gives the right to the owner, including it in the bundle labelled 'ownership of land', it will be difficult for the parents to buy it, since the parents face a public good problem in purchasing an agreement from the owner to put high fences around his tanks. Hence we have an argument for the existing law of attractive nuisance, under which the parent can enjoin the property owner from leaving the tanks unfenced, or sue for damages if his child is injured. This is one example of the way in which the Coase Theorem approach helps illuminate a wide range of legal issues.

Prediction: What the Law Will Be

Economic analysis, of law or anything else, can be viewed either as an attempt to learn what should be or as an attempt to explain what is and predict what will be. In the case of the economic analysis of the law, attempts to explain and predict have taken two rather different forms.

On the one hand, there is the argument of Richard Posner, according to which the common law tends, for a variety of reasons, to be economically efficient. The analysis of what legal rules are efficient thus provides an explanation of what legal rules exist – and the observation of what legal rules exist provides a test of theories about what rules are efficient.

On the other hand, there is the approach associated with public choice theory, which views legislated, administrative, and perhaps even common law as outcomes of a political market on which interest groups seek private objectives by governmental means. Since the amount a group is willing to spend in order to get the laws it favours depends not only on the value of the law to that group but also on the group's ability to solve the public good problem of inducing its members to contribute, expenditures in the political market will not accurately represent the value of the law to those affected, hence inefficient laws – laws which injure the losers by more than they benefit the gainers – may well pass, and efficient laws may well fail. The most obvious implication of this line of analysis is that laws will tend to favour concentrated interests at the expense of dispersed interests, since the former will be better able to raise money from their members to lobby for the laws they prefer.

Conclusions

In looking at economic analysis of law, one striking observation is the way in which economists tend to convert issues from disputes about equity, justice, fairness or the like into disputes about efficiency. In part, this is because economists do, and traditional legal scholars often do not, take account of the effect of legal rules on market

prices. The result of taking these effects into account is frequently to eliminate the distributional effects of changes in such rules. In part, it is because economists do, and legal scholars sometimes do not, assume that rules modify behaviour. If so, then in evaluating the rules we must ask not only whether they produce a just outcome in a particular case, but whether their effects on the behaviour of those who know of the rules and modify their actions to take account of them is in some sense desirable.

A second observation is that economic analysis frequently demonstrates the existence of efficiency arguments for rules usually thought of as based entirely on considerations of justice. One simple example is the law against theft. At first glance, theft appears to involve no question of economic efficiency at all; the thief is better off by the same amount by which the victim is worse off, hence the transaction, however unjust, is not inefficient.

That conclusion is wrong. The opportunity to gain by stealing diverts resources to that activity. In equilibrium, the marginal thief receives the same income from stealing (net of risk of imprisonment, cost of tools, etc.) as he would in some alternative productive activity; there is no gain to the marginal thief to balance the cost to the victim. Hence theft can be condemned as inefficient with no reference to issues of justice.

A third observation is the degree to which the examination of real legal issues and real cases forces the economist to take account of some of the complexities of real-world interactions which he might otherwise never notice, and thus provides him with the opportunity to increase the depth and power of his analysis.

A final, and important, observation is that economics provides a unity among disparate fields of law which is lacking in much traditional legal analysis. In the words of one of the field's leading practitioners:

> Almost any tort problem can be solved as a contract problem, by asking what the people involved in an accident would have agreed on in advance with regard to safety measures if transaction costs had not been prohibitive . . ..Equally, almost any contract problem can be solved as a tort problem by asking what sanction is necessary to prevent the performing or paying party from engaging in wasteful conduct, such as taking advantage of the vulnerability of a party who performs his side of the bargain first. And both tort and contract problems can be framed as problems in the definition of property rights; for example, the law of negligence could be thought to define the right we have in the safety of our persons against accidental injury. The definition of property rights can itself be viewed as a process of figuring out what measures parties would agree to, if transaction costs weren't prohibitive, in order to create incentives to avoid wasting valuable resources (Posner 1986).

Any note as short as this can provide only a very incomplete description of the field, and one heavily biased towards the author's own interests. The references cited below, and the references in Posner (1986) and Goetz (1984), provide a much more extensive survey.

See Also

- Coase Theorem
- Common law
- Crime and punishment
- Natural law
- Property rights

Bibliography

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76(March): 169–217.

Becker, G. 1976. *The economic approach to human behavior.* Chicago: University of Chicago Press.

Calabresi, G. 1961. Some thoughts on risk distribution and the law of torts. *Yale Law Journal* 70(March): 499–553.

Calabresi, G., and A.D. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathederal. *Harvard Law Review* 85(6): 1089–1182.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3(October): 1–44.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review, Papers and Proceedings* 57(2): 347–359, especially 351–353.

Ehrlich, I. 1972. The deterrent effect of criminal law enforcement. *Journal of Legal Studies* 1(2): 259–276.

Goetz, C.J. 1984. *Cases and materials on law and economics*. St Paul: West.

Landes, E.M. 1982. Insurance, liability, and accidents: A theoretical and empirical investigation of the effect

of no-fault on accidents. *Journal of Law and Economics* 25(1): 49–65.
Landes, W., and R. Posner. 1978. Salvors, finders, good samaritans, and other rescuers: An economic study of law and altruism. *Journal of Legal Studies* 7(1): 83–128.
McGee, J.S. 1958. Predatory price cutting: The Standard Oil (N.J.) case. *Journal of Law and Economics* 1-(October): 137–169.
Peltzman, S. 1975. The effects of automobile safety regulations. *Journal of Political Economy* 83(4): 677–725.
Posner, R. 1986. *Economic analysis of law*. Boston: Little, Brown.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: W. Strahan & T. Cadell.
Tullock, G. 1971. *The logic of the law*. New York: Basic Books.

Law and Economics of Copyright and Trademarks on the Internet

Stefan Bechtold

This article provides an overview of the new opportunities and challenges for copyright and trademark law and economics research created by the Internet. It reviews the relevant literature in the field of copyright law, in particular as it relates to piracy, liability and market structure, Digital Rights Management, the relationship between copyright and contract, and incentives for creativity. As far as trademark law is concerned, the article describes the empirical literature, focusing on keyword advertising. The article concludes with an outlook on future research areas and methodologies.

Introduction

The rise of the Internet and other new communication technologies has created novel challenges for copyright and trademark law. On the one hand, courts worldwide are being called upon to deal with problems of liability in peer-to-peer file-sharing networks and in keyword advertising. On the other hand, new technological developments have enabled scholars to subject some of the basic assumptions of copyright and trademark protection to fresh scrutiny. The increasing participation of Internet users in the creation, selection and distribution of information has raised questions such as how far copyright is enforceable or even necessary in an online world, and how novel uses of brands on the Internet affect traditional trademark doctrines.

This review focuses on how copyright and trademark law and economics research cope with new technological developments and whether these developments alter the basic structure of copyright and trademark law. It puts emphasis on the literature at the intersection of law and economics. Purely economic and legal research, and research into the law and economics of copyright and trademarks outside the Internet, will be covered only to the extent that is necessary for the review. For more general reviews of law and economics research into copyright and trademark law both on- and offline, the reader is directed to Landes and Posner (2003), Menell and Scotchmer (2007), Burk (2012) and Handke (2012).

Copyright Protection

For a long time, questions of copyright policy were of limited interest outside certain well-defined industries and academic circles. With the advent of the Internet, copyright policy started to make headlines on a continuous basis, and this led to a broad range of research. Law and economics research on copyright on the Internet has focused on five areas in particular: (1) piracy, (2) liability and market structure, (3) Digital Rights Management, (4) the relationship between copyright and contract, and (5) incentives to create, contribute and distribute.

Piracy

Digital technologies have enabled users to reproduce content with unprecedented speed and ease. Peer-to-peer (P2P) file-sharing networks have enabled Internet users to distribute and download

an astonishing amount of copyrighted works – be they movies, songs or other content – without authorisation from copyright owners. At the same time, every reproduction of a copyrighted work made by a digital device may potentially lead to a copyright infringement, unless some copyright exception applies (such as the US fair use defence or Article 5 of the EU Copyright Directive 2001). This has led to abundant discussion in the theoretical and empirical economics literature on the relationship between copyright protection and piracy on the Internet.

The theoretical literature is expansive. For more extensive reviews of this area, see Peitz and Waelbroeck (2006a), as well as Belleflamme and Peitz (2012). In general, the literature focuses on three areas. First, building upon earlier literature on the effects of photocopying on copyright protection in general (Novos and Waldman 1984; Liebowitz 1985; Johnson 1985; Besen and Kirby 1989; Landes and Posner 1989), a new generation of economics researchers have analysed the effects of online piracy on industry profits and welfare. Yoon (2002) and Banerjee (2003) analyse the optimal level of copyright protection and point to the divergent interests of content producers and society at large. Cho and Ahn (2010) study the way in which piracy affects the offering of content in different versions and the impact that such product differentiation has on copyright protection. Wu and Chen (2008) analyse the extent to which versioning can be used to combat piracy. Bae and Choi (2006) introduce a model with vertical product differentiation in which unauthorised copies are of lower quality.

A second strand of the theoretical literature looks at the potential benefits to content producers of tolerating piracy. This literature suggests reasons why firms may have an incentive not to use the strongest legal and technical means of protection available. Some of this literature may also support policy arguments for weaker copyright protection. Gopal et al. (2006), Duchêne and Waelbroeck (2006) and Peitz and Waelbroeck (2006b) analyse the extent to which unauthorised file sharing can increase the attractiveness of the original product, as it enables consumers to find out whether they like the product before making a purchase decision (sampling effect). Conner and Rumelt (1991), Takeyama (1994), Shy and Thisse (1999), Gayer and Shy (2003b, 2006), Jain (2008) and Herings et al. (2010) analyse whether network effects, which arise not only from authorised but also from unauthorised users, can increase the attractiveness of copyrighted works for all users. King and Lampe (2003) point out that benefits due to network effects may disappear if the content producer engages in versioning strategies. Rasch and Wenzel (2013) extend the network effects analysis to a two-sided market setting of a software platform provider. Turning to monopoly analysis, Martínez-Sánchez (2010) studies how the presence of unauthorised copies can help a social planner to reduce the negative social welfare effects of a monopolist. Concerning externalities between users, August and Tunca (2008), and Lahiri (2012) analyse whether software vendors should allow users of unauthorised software copies to apply security patches.

A third strand of the theoretical literature analyses litigation and enforcement strategies in a world of massive online piracy. Harbaugh and Khemka (2010), as well as Cremer and Pestieau (2009), analyse enforcement strategies which target high-valuation consumers only. Takeyama (2009) shows that a content producer may signal product quality through its copyright enforcement decision. Lahiri and Dey (2013) analyse whether lower piracy enforcement can increase a content producer's incentive to invest in product quality.

The empirical literature on the relationship between online piracy and copyright protection does not lag behind its theoretical counterpart as far as breadth and scope are concerned. The focus of this literature has been to identify the impact that online piracy has on industry profits. The music industry, for example, claims that the emergence of peer-to-peer file-sharing networks and the resulting massive online pirating of music has been a major factor contributing to the decrease in sales of physical music recordings in various countries. Using econometric methods to provide evidence for the existence and magnitude of this effect is a complex task, for two reasons. First, illegal behaviour is neither readily documented nor easily observable on the Internet.

Second, as in other areas of law and economics, it is methodologically challenging to establish a causal relationship between piracy and industry profits using standard methods, owing to a number of endogeneity concerns. As a consequence, the results of the empirical studies depend on the data and methods used. For an extensive review of this literature and a discussion of the methodological challenges, see Waldfogel (2012c) and Handke (2012); see also Waldfogel (2012d), Oberholzer-Gee and Strumpf (2010) and Liebowitz (2006).

Most papers in the field analyse whether the increase in online piracy has led to a decrease in sales of physical music recordings (displacement effect). Most prominently, Oberholzer-Gee and Strumpf (2007) find no such displacement effect in their data. They match a sample of downloads from peer-to-peer file-sharing networks to US sales data for a large number of albums. A potential policy conclusion is that file sharing cannot be regarded as the primary reason for the decline in music sales. In a related vein, Bhattacharjee et al. (2007) find a negative effect for lower debut ranked albums, but no significant effect for top debut ranked albums.

Other papers find a displacement effect. They use consumer surveys to identify changes in consumer behaviour, or use countries, cities or records as unit of analysis. Using various data sources and methods, Peitz and Waelbroeck (2004), Zentner (2006), Rob and Waldfogel (2006), Michel (2006), and Liebowitz (2008) and Barker (2012) find that file sharing makes people less likely to buy music recordings. Waldfogel (2010) arrives at similar results, focusing on a time in which legitimate digital alternatives (iTunes) were already available. Andersen and Frenz (2010) point to the countervailing effect that consumers may use file-sharing networks as part of their purchase decision process, trying out music before buying the product (sampling). While most studies focus on the music industry, some papers study displacement effects in the movie industry, again with mixed results as regards the existence and size of the displacement effect; see Smith and Telang (2009), Bounie et al. (2006), Rob and Waldfogel (2007), Zentner (2010), Bai and Waldfogel (2012) and Danaher et al. (2010).

One shortcoming of many displacement effect studies is that they focus on industry profits only, taking no account of wider effects on social welfare (but see Rob and Waldfogel 2006). Another shortcoming is that they can draw conclusions only about the quantity of content produced and consumed, not about the quality. In this respect, Waldfogel (2012a) analyses whether the emergence of file-sharing networks can be linked to a decrease in the quality of music being created. He finds no evidence of a quality reduction.

Beyond the displacement effect, empirical literature analyses the impact of enforcement and litigation strategies on Internet user behaviour. In the early 2000s, the recording industry in the USA and Europe started to take highly publicised legal action against individual file sharers, demanding very high (statutory) damages from individuals. While the goal was to achieve maximum general deterrence by focusing on a small number of individual high-volume users, the success of this mass-litigation strategy seems mixed. By tracking file sharing activity on peer-to-peer file-sharing networks, Bhattacharjee et al. (2006) find that, while the mass-litigation actions substantially reduced the level of file-sharing activity, unauthorised content remained easily available on the networks. After a period of mass litigation, various countries enacted or considered enacting a so-called 'three-strikes law'. Such a law empowers an administrative authority to send a graduated system of warnings to identified copyright infringers. If the infringers do not comply, this can lead to the temporary suspension of their Internet access. Following the enactment of such a law in France, Danaher et al. (2013) analyse the impact of increased law enforcement on Internet user behaviour. Depoorter et al. (2011) point to an interaction between the level of copyright enforcement, deterrence and social norms: where copyright infringements are widespread in a society, enforcement may have to be raised to a level which undermines the society's support for the underlying copyright rules. Balestrino (2008) presents a model of social norm formation to explain why Internet users fail to adhere to copyright laws

while remaining lawabiding citizens in other areas of social life.

Liability and Market Structure

While the economics-oriented literature has concentrated on identifying and measuring the impact of online piracy, the more law-oriented literature has analysed copyright liability and its relationship to market structure over the last few years. The literature focuses on the liability of distribution technology providers, the status of other intermediaries and proposals to broaden the scope of liability rule regimes in copyright law.

With respect to the liability of distribution technology providers, the commercialisation of the Internet in the early 1990s has led to significant changes in value chains. Traditionally strong intermediaries – such as record companies, publishers, newspapers and movie companies – have been struggling to define their future roles and find profitable business models in a radically changed environment of content consumption. At the same time, new intermediaries – such as search engines, auction sites and social networks – have emerged. In some cases, they threaten to displace traditional intermediaries. In other cases, they complement them or create entirely new business models. Defining their legal responsibilities is an important aspect of digital copyright policy.

Particularly important is the specific design of property entitlements. The stronger copyright protection becomes in the digital world, the better copyright owners may control new technological uses of their works, ideally leading to greater *ex ante* incentives to produce such works. At the same time, allocating property rights to creators may impede the development and deployment of new distribution technologies, whose developers bear a higher liability risk under this entitlement regime. The tradeoff between providing incentives to digital content producers and providing them to distribution technology developers was already at the core of the U.S. Supreme Court decision in the Betamax case (Sony Corp. of America v. Universal City Studios, 464 U.S. 417 (1984)). It gained new importance with the growth in peer-to-peer file-sharing technologies (MGM Studios v. Grokster, 545 U.S. 913 (2005)).

A series of articles explores this tradeoff, which may be affected by the design of the copyright and liability regime. Wu (2005) and Barnett (2013) make the general point that copyright policy is not only about providing incentives to creators. It is also an industrial policy regulating competition among rival distribution technology providers. Landes and Lichtman (2003) as well as Lichtman and Landes (2003) point out that distribution technology providers are often in a good position to monitor direct copyright infringers or to redesign their distribution systems so as to make direct infringement more difficult. They weigh these arguments for expanding secondary liability against the effects that such expansion may have on the legitimate use of relevant tools, services and venues (dual use problem) (see also Menell 2009). Oliar (2012) analyses this tradeoff from a property and liability rules perspective (see Calabresi and Melamed 1972). Oliar proposes a modifiable entitlement regime that maximises innovation incentives for both content creators and distribution technology developers while minimising investment distortions. Lemley and Reese (2004) point to the socially harmful consequences of expanding secondary liability. They explore three alternative enforcement strategies: increasing deterrence for copyright infringers by raising the cost of direct infringement (by introducing criminal sanctions or increasing monetary damages); reducing the costs of enforcement against individual infringers (by introducing a levy system); or introducing an effective dispute resolution system.

As regards the status of copyright intermediaries, the relationship between digital copyright policy and market structure has been discussed in the context of the Google Books project. Since Google announced its plans to digitise millions of books and make them available worldwide without express opt-in authorisation from the respective copyright owners, the project has led to heated controversies around the globe. Bracha (2007) shows that the debate about whether copyright law should move from an opt-in to an opt-out system in the context of Google Books and related orphan works problems is a good example of a *Hohfeldian* rearrangement of property rights.

Müller-Langer and Scheufen (2011) view the Google Books project as a potential solution to the orphan works problem. Lichtman (2009) points out that, with respect to Google Books, copyright can be used to organise and structure distribution technology markets. In these markets, copyright protection creates an entry barrier for novel intermediaries whose negative effects have to be weighed against the positive effects of copyright protection on authors and their traditional intermediaries.

In addition to Google Books, the status of copyright intermediaries is also discussed with regard to technology platforms such as video game platforms, cell phones and tablet devices. This literature is both theoretical (Lichtman 2000; Economides and Katsamakas 2006) and empirical (Boudreau 2010). As the literature usually deals with intellectual property protection in general as opposed to copyright protection in particular, the reader is referred to Armstrong and Wright (2008) for more information. On a slightly different matter, Samuelson and Scotchmer (2002) analyse the relationship between copyright policy and market structure with regard to reverse engineering in the software industry and to content protected by Digital Rights Management.

The relationship between copyright policy and market structure also lies at the core of empirical analyses of information aggregators. In a study of aggregated website log data, Chiou and Tucker (2011) analyse how Internet users are using information aggregators such as Google News. They find that news aggregators do not serve as a complete substitute for the underlying websites. Rather, they encourage users to navigate further. Such empirical findings have potentially important policy consequences when it comes to determining the liability of information aggregators and the harm they may do to copyright owners.

Copyright law can also affect market structure by introducing or expanding liability rule regimes. For several decades, copyright law has been grappling with the question of how to deal with private copying by consumers. With the emergence of cassette recorders and photocopying machines, it became evident that private copying was a mass phenomenon that was very hard to control. For this and other reasons, many European countries introduced a copyright exemption for private copying, but compensated rights holders indirectly by a levy system which imposed a levy on all blank media and copying devices being sold. In the USA, the Audio Home Recording Act of 1992 is the only piece of legislation to include a levy system (for digital audio recording devices and blank storage media). While US scholars have proposed a considerable expansion of levy systems in the digital environment in order to cope with digital piracy (Netanel 2003; Fisher 2004), European scholars have typically remained more sceptical, given the experience with levy systems in Europe (Bechtold 2003). Chen and Png (2003) provide an analytical framework suggesting how governments should use available policy instruments – penalties, levy systems, subsidies – for digital content. Gayer and Shy (2003a) identify circumstances in which levy systems taxing hardware equipment are inefficient. Alcalá and González-Maestre (2010) point to the potentially different effects of a levy system on superstars as compared to young artists.

Digital Rights Management

While digital technologies have enabled large-scale piracy, they may also provide a solution to the problem. 'Digital Rights Management' (DRM) promises a secure framework for distributing digital content, ensuring that rights holders receive adequate remuneration for the creation of their content. Compared to traditional copyright law, an ideal DRM system promises an unprecedented degree of control over the entire distribution chain, and the usage of digital content, by combining different means of protection, in particular technology, contracts, technology licenses, anti-circumvention regulations and traditional copyright protection. Although adoption of DRM systems has been significantly lower than some predicted in the late 1990s, DRM technology remains an important building block of protection for various distribution technologies, such as DVDs, mobile communications and non-PC handheld devices.

One strand of the literature focuses on DRM as an alternative to copyright protection, potentially

overriding limitations to copyright protection such as the US fair use defence (Bechtold 2003, 2004; Elkin-Koren and Salzberger 2013). It also points to shortcomings of the legal literature in so far as it applies simplistic industrial organisation models to DRM (Cohen 1998).

The more economics-oriented DRM literature focuses on the use of DRM to set the level of protection endogenously through technology. Increasing protection makes copyright infringement more costly, but potentially decreases the value of protected content for lawful users. Ahn and Shin (2010) analyse the optimal level of DRM protection by endogenising protection levels and taking account of supply-side substitution effects in copyright enforcement. Sundararajan (2004) offers a related model in which the presence of DRM decreases the valuation of both authorised and unauthorised products. In his model, the maximum level of DRM protection is optimal for a content provider in the absence of price discrimination. Vernik et al. (2011) focus on demand-side substitution effects between protected content and unauthorised file-sharing. In their model, download piracy decreases when a firm offers authorised DRM-free downloads, and company profits do not necessarily increase as it becomes harder to engage in piracy. Park and Scotchmer (2005) analyse the effect of DRM on equilibrium price in an oligopoly setting where firms can decide whether or not to share DRM technologies. Choi et al. (2010) analyse how the strategic interaction among content producers affects their choice of a particular level of DRM protection.

Relationship Between Copyright and Contract

Following a decision by the Seventh Circuit Court of Appeals in the USA (ProCD, Inc. v. Zeidenberg, 86 F.3d. 1147 (7th Cir. 1996)), legal scholars have discussed extensively whether limitations to copyright protection, such as the US fair use defence, can be waived by contract, particularly in a mass-market context. Such contracts, which are offered on a take-it-or-leave-it basis, impose standardised terms on large numbers of customers, leaving them unable to influence the specific contractual terms. While some courts and scholars argue that no legal intervention is necessary because competition among vendors ensures adequate consumer protection in such a market (see ProCD, Inc. v. Zeidenberg, 86 F.3d 1147, 1453 (7th Cir. 1996)), other scholars have pointed to various market failures that impede such competition (Elkin-Koren 1997; Merges 1997). In the USA, this question primarily concerns the relationship between federal copyright and state contract law and its governing preemption doctrine (Lemley 1995). In Europe, the question is covered by scattered provisions in various European copyright directives and relates to the relationship between European copyright and contract law (Guibault 2002).

Incentives to Create, Contribute and Distribute

The Internet has opened up new ways to engage in creative and collaborative activities. It has contributed to the flourishing of various 'open innovation' paradigms, in particular 'open source' and 'open access'. The economics of open source software is covered by a separate review article, to which the reader is referred (Fershtman and Gandal 2011). More generally, new communication technologies have posed the question of how important it is to provide potential creators with extrinsic incentives when creative output is highly modular and intrinsically motivated (Benkler 2002). Jian and Mackie-Mason (2012) point to alternative ways, apart from copyright protection, to incentivise the production and quality screening of usergenerated content. Brynjolfsson and Zhang (2006) propose a 'statistical couponing mechanism' that should incentivise the creation of digital goods while producing a significantly lower deadweight loss compared to traditional intellectual property protection.

The development of new approaches to the incentive paradigm in copyright law is not restricted to creators, but extends to users as well. Strahilevitz (2003), Nandi and Rochelandet (2008) and Casadesus-Masanell and Hervas-Drane (2010) provide rational-choice and behavioural explanations for the willingness of Internet users to upload content and share bandwidth on file-sharing networks. Regner and Barria

(2009) provide an empirical analysis of voluntary contributions to an online music label.

Trademark

While the law and economics literature relating to copyright issues on the Internet is enormous, the corresponding trademark literature is at a much earlier stage. A large industrial organisation literature deals with issues of advertising and branding, but rarely focuses on trademark law in general, or trademark law on the Internet in particular. At the same time, a large legal trademark literature has responded to various waves of technological development, focusing particularly on hyperlinking, meta-tags and domain names, but has rarely adopted a distinct law and economics perspective.

It was the emergence of keyword advertising in search engines that finally led to research into Internet trademark issues that really lies on the borderline between law and economics. Internet search engines display advertisements along with search results, which constitute a major source of revenue for the search engines. The display of ads is triggered by the use of keywords. If an advertiser buys a keyword which contains a trademark owned by another company, the trademark owner often tries to prevent its rival and the search engine operator from using its trademark in such a way. The extent to which advertisers and – through either primary or secondary liability doctrines – search engine operators are liable for trademark infringement with regard to keyword advertising is vigorously debated in the USA, Europe and beyond.

From a trademark law perspective, it is questionable whether a search engine operator "uses" a keyword as a trademark, and whether consumers are likely to be confused by such use. Legal scholars have remained skeptical of such expansive interpretations of trademark protection (Dogan and Lemley 2004; Goldman 2005), but the empirical basis of their arguments has been limited. Increasingly, scholarly attention is focusing on empirical investigations of the relationship between trademark policies and keyword advertising. Some empirical studies use changes in keyword advertising policies of search engine operators to measure the impact of the policy change on user behaviour. Chiou and Tucker (2012) draw on aggregate use search data to show that allowing resellers to use third-party trademarks as keywords without the third party's authorisation reduces the number of clicks on the trademark owner's paid search ads. However, this is outweighed by a large increase in consumers clicking on the unpaid links to the trademark owner's website within the main search results (substitution effect between paid search results and main search results).

Bechtold and Tucker (2013) use individual user-level click-stream data to analyse how the search behaviour of Internet users in Germany and France changed after Google allowed third parties in Europe to register trademarks as keywords in 2010. They find opposing effects: while navigational searches are less likely to lead to the trademark owner's website, non-navigational searches are more likely to lead to the trademark owner's website after the policy change. This indicates that, in a keyword advertising system in which control rights over keyword advertising are fully allocated to trademark owners, the positive effects on trademark owners and some search engine users may, potentially, be counterweighed by negative effects on other users and also on trademark owners. Franklyn and Hyman (2013) use a combination of consumer surveys and an analysis of actual advertisements displayed by search engines. They find little evidence or risk of consumer confusion and a high degree of willingness by consumers to purchase competing products. Similarly, by analysing actual advertisements displayed by search engines on a smaller scale, Rosso and Jansen (2010) question whether third-party keyword advertising is a widespread phenomenon. In general, these studies demonstrate that Internet users are using trademarks in much more subtle and varied ways than is often assumed in the trademark discourse.

Future Research

While the application of law and economics research methodologies has led to a burgeoning

of insightful social science research on issues of digital copyright and trademark policy, much work remains to be done. From a methodological perspective, future law and economics research on copyright and trademark on the Internet will follow the general trend in law and economics. Internet research, in particular, will increasingly be based on empirical methods. This will include econometric analyses, experimental studies in the laboratory and field experiments. The Internet provides an unprecedented amount of data for such studies (Edelman 2012).

From a legal perspective, various areas of digital copyright and trademark policy have not been thoroughly analysed by law and economics researchers. This includes institutional questions such as the role of copyright collecting societies in the digital environment (but see Katz (2006) as well as Handke and Towse (2007)); a focus on data and institutional arrangements outside the USA (e.g. on the role of moral rights on the Internet in European copyright systems); an analysis of political economy dimensions (see Banerjee 2006); and the integration of behavioural approaches into the analysis. The analysis of copyright issues is further developed than trademark discussions.

From an economics perspective, research is impeded by the limited availability of data in the area of copyright, while rigorous mining of available trademark registration and litigation data is only just beginning. Another challenge for economics research is to go beyond established research paradigms. Such a move could contribute substantially to the further integration of law and economics research in the field if, for example, empirical studies were to focus not only on right owners' revenues, but also on social welfare. Another example is the attempt to analyse not only how digitisation has affected the quantity but also the quality of works being produced and how this has affected consumer product discovery (Waldfogel 2012b).

See Also

- ▶ Computer Industry
- ▶ Electronic Commerce
- ▶ Intellectual Property
- ▶ Internet and the Offline World
- ▶ Internet, Economics of the
- ▶ Music Markets, Economics of
- ▶ Network Goods (Empirical Studies)
- ▶ Online Platforms, Economics of
- ▶ Open Source Software, A Brief Survey of the Economics of
- ▶ Two-Sided Markets

Acknowledgment The author would like to thank Aurelia Tamo` for very helpful research assistance.

Bibliography

Ahn, I., and I. Shin. 2010. On the optimal level of protection in DRM. *Information Economics and Policy* 22: 341–353.

Alcalá, F., and M. González-Maestre. 2010. Copying, superstars, and artistic creation. *Information Economics and Policy* 22: 365–378.

Andersen, B., and M. Frenz. 2010. Don't blame the P2P file-sharers: The impact of free music downloads on the purchase of music CDs in canada. *Journal of Evolutionary Economics* 20: 715–740.

Armstrong, M. and J. Wright. 2008. Two-sided markets. In *The new Palgrave dictionary of economics*, 2nd edn, ed. S. Durlauf and L. Blume. Online edition.

August, T., and T.I. Tunca. 2008. Let the pirates patch? An economic analysis of software security patch restrictions. *Information Systems Research* 19: 48–70.

Bae, S.H., and J.P. Choi. 2006. A model of piracy. *Information Economics and Policy* 18: 303–320.

Bai, J., and J. Waldfogel. 2012. Movie piracy and sales displacement in two samples of Chinese consumers. *Information Economics and Policy* 24: 187–196.

Balestrino, A. 2008. It is a theft but not a crime. *European Journal of Political Economy* 24: 455–469.

Banerjee, D.S. 2003. Software piracy: A strategic analysis and policy instruments. *International Journal of Industrial Organization* 21: 97–127.

Banerjee, D.S. 2006. Lobbying and commercial software piracy. *European Journal of Political Economy* 22: 139–155.

Barker, G. 2012. Evidence of the effect of free music downloads on the purchase of music CDs in Canada. *Review of Economic Research on Copyright Issues* 9: 55–78.

Barnett, J.M. 2013. *Copyright without creators*. Manuscript available at http://ssrn.com/abstract=2245038. Accessed 9 May 2013.

Bechtold, S. 2003. The present and future of digital rights management: Musings on emerging legal problems. In *Digital rights management: Technological, economic, legal and political aspects*, ed. E. Becker, W. Buhse, D. Gu¨ nnewig, and N. Rump, 597–654. Berlin: Springer.

Bechtold, S. 2004. Digital rights management in the United States and Europe. *American Journal of Comparative Law* 52: 323–382.

Bechtold, S., and C. Tucker. 2013. *Trademarks, Triggers and online search*. Manuscript available at http://ssrn.com/abstract=2266945. Accessed 20 May 2013.

Belleflamme, P., and M. Peitz. 2012. Digital piracy: Theory. In *The oxford handbook of the digital economy*, ed. M. Peitz and J. Waldfogel, 489–530. Oxford: Oxford University Press.

Benkler, Y. 2002. Coase's penguin, or, Linux and the nature of the firm. *Yale Law Journal* 112: 369–446.

Besen, S.M., and S.N. Kirby. 1989. Private copying, appropriability, and optimal copying royalties. *Journal of Law & Economics* 32: 255–280.

Bhattacharjee, S., R.D. Gopal, K. Lertwachara, and J.R. Marsden. 2006. Impact of legal threats on online music sharing activity: An analysis of music industry legal actions. *Journal of Law & Economics* 49: 91–114.

Bhattacharjee, S., R.D. Gopal, K. Lertwachara, J.R. Marsden, and R. Telang. 2007. The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science* 53: 1359–1374.

Boudreau, K.J. 2010. Open platform strategies and innovation: Granting access vs. Devolving control. *Management Science* 56: 1849–1872.

Bounie, D., M. Bourreau, and P. Waelbroeck. 2006. Piracy and the demand for films: Analysis of piracy behavior in French universities. *Review of Economic Research on Copyright Issues* 3: 15–27.

Bracha, O. 2007. Standing copyright law on its head? The Googlization of everything and the many faces of property. *Texas Law Review* 85: 1799–1869.

Brynjolfsson, E., and X. Zhang. 2006. Innovation incentives for information goods. In *Information policy and the economy*, vol. 7, ed. J. Lerner and S. Stern, 99–123. Cambridge: MIT Press.

Burk, D.L. 2012. Law and economics of intellectual property: In search of first principles. *Annual Review of Law and Social Science* 8: 397–414.

Calabresi, G., and A.D. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review* 85: 1089–1128.

Casadesus-Masanell, R., and A. Hervas-Drane. 2010. Peer-to-peer file sharing and the market for digital information goods. *Journal of Economics & Management Strategy* 19: 333–373.

Chen, Y.-N., and I. Png. 2003. Information goods pricing and copyright enforcement: Welfare analysis. *Information Systems Research* 14: 107–123.

Chiou, L., and C. Tucker. 2011. *Copyright, digitization, and aggregation*. Manuscript available at http://ssrn.com/abstract=1864203. Accessed 9 May 2013.

Chiou, L., and C. Tucker. 2012. How does the use of trademarks by third-party sellers affect online search? *Marketing Science* 31: 819–837.

Cho, W.-Y., and B.-H. Ahn. 2010. Versioning of information goods under the threat of piracy. *Information Economics and Policy* 22: 332–340.

Choi, P., S.H. Bae, and J. Jun. 2010. Digital piracy and firms' strategic interactions: The effects of public copy protection and DRM similarity. *Information Economics and Policy* 22: 354–364.

Cohen, J.E. 1998. Lochner in cyberspace: The new economic orthodoxy of 'rights management'. *Michigan Law Review* 97: 462–564.

Conner, K.R., and R.P. Rumelt. 1991. Software piracy: An analysis of protection strategies. *Management Science* 37: 125–139.

Cremer, H., and P. Pestieau. 2009. Piracy prevention and the pricing of information goods. *Information Economics and Policy* 21: 34–42.

Danaher, B., S. Dhanasobhon, M.D. Smith, and R. Telang. 2010. Converting pirates without cannibalizing purchasers: The impact of digital distribution on physical sales and Internet piracy. *Marketing Science* 29: 1138–1151.

Danaher, B., M.D. Smith, R. Telang, and S. Chen. 2013. The effect of graduated response anti-piracy laws on music sales: Evidence from an event study in France. *Journal of Industrial Economics* 62: 541–553.

Depoorter, B., A.v. Hiel, and S. Vanneste. 2011. Copyright backlash. *Southern California Law Review* 84: 1251–1292.

Dogan, S.L., and M.A. Lemley. 2004. Trademarks and consumer search costs on the Internet. *Houston Law Review* 41: 777–838.

Duchêne, A., and P. Waelbroeck. 2006. The legal and technological battle in the music industry: Information-push versus information-pull technologies. *International Review of Law and Economics* 26: 565–580.

Economides, N., and E. Katsamakas. 2006. Two-sided competition of proprietary vs. open source technology platforms and the implications for the software industry. *Management Science* 52: 1057–1071.

Edelman, B. 2012. Using Internet data for economic research. *Journal of Economic Perspectives* 26: 189–206.

Elkin-Koren, N. 1997. Copyright policy and the limits of freedom of contract. *Berkeley Technology Law Journal* 12: 93–113.

Elkin-Koren, N., and E.M. Salzberger. 2013. *The law and economics of intellectual property in the digital age: The limits of analysis*. London: Routledge.

EU Copyright Directive. 2001. Directive 2001/29/EC of the European parliament and of the council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal of the European Communities*, L 167, 22 June 2001, pp. 10–19.

Fershtman, C. and N. Gandal. 2011. A brief survey of the economics of open source software. In: *The new Palgrave dictionary of economics*, 2nd edn, ed. S. Durlauf and L. Blume. Online edition.

Fisher, W.W. 2004. *Promises to keep: Technology, law, and the future of entertainment*. Stanford: Stanford University Press.

Franklyn, D. J. and D. A. Hyman. 2013. Trademarks as keywords: Much ado about something?. *Harvard Journal of Law & Technology* (in press).

Gayer, A., and O. Shy. 2003a. Copyright protection and hardware taxation. *Information Economics and Policy* 15: 467–483.

Gayer, A., and O. Shy. 2003b. Internet and peer-to-peer distributions in markets for digital products. *Economics Letters* 81: 197–203.

Gayer, A., and O. Shy. 2006. Copyright enforcement in the digital era. In *Industrial organization and the digital economy*, ed. G. Illing and M. Peitz, 229–240. Cambridge: MIT Press.

Goldman, E. 2005. Deregulating relevancy in Internet trademark law. *Emory Law Journal* 54: 507–596.

Gopal, R.D., S. Bhattacharjee, and G.L. Sanders. 2006. Do artists benefit from online music sharing? *Journal of Business* 79: 1503–1533.

Guibault, L. 2002. *Copyright limitations and contracts: An analysis of the contractual overridability of limitations on copyright*. London: Kluwer Law International.

Handke, C. 2012. A taxonomy of empirical research on copyright: How do we inform policy? *Review of Economic Research on Copyright Issues* 9: 47–92.

Handke, C., and R. Towse. 2007. Economics of copyright collecting societies. *International Review of Intellectual Property & Competition Law* 8: 937–957.

Harbaugh, R., and R. Khemka. 2010. Does copyright enforcement encourage piracy? *Journal of Industrial Economics* 58: 306–323.

Herings, J.-J.P., R. Peeters, and M.S. Yang. 2010. Competition against peer-to-peer networks. *Information Economics and Policy* 22: 315–331.

Jain, S. 2008. Digital piracy: A competitive analysis. *Marketing Science* 27: 610–626.

Jian, L., and J.K. Mackie-Mason. 2012. Incentive-centered design for usercontributed content. In *The oxford handbook of the digital economy*, ed. M. Peitz and J. Waldfogel, 399–433. Oxford: Oxford University Press.

Johnson, W.R. 1985. The economics of copying. *Journal of Political Economy* 93: 158–174.

Katz, A. 2006. The potential demise of another natural monopoly: New technologies and the administration of performing rights. *Journal of Competition Law & Economics* 2: 245–284.

King, S.P., and R. Lampe. 2003. Network externalities, price discrimination and profitable piracy. *Information Economics and Policy* 15: 271–290.

Lahiri, A. 2012. Revisiting the incentive to tolerate illegal distribution of software products. *Decision Support Systems* 53: 357–367.

Lahiri, A., and D. Dey. 2013. Effects of piracy on quality of information goods. *Management Science* 59: 245–264.

Landes, W.M., and D. Lichtman. 2003. Indirect liability for copyright infringement: Napster and beyond. *Journal of Economic Perspectives* 17: 113–124.

Landes, W.M., and R.A. Posner. 1989. An economic analysis of copyright law. *Journal of Legal Studies* 18: 325–364.

Landes, W.M., and R.A. Posner. 2003. *The economic structure of intellectual property law*. Cambridge: Harvard University Press.

Lemley, M.A. 1995. Intellectual property and shrinkwrap licenses. *Southern California Law Review* 68: 1239–1294.

Lemley, M.A., and R.A. Reese. 2004. Reducing digital copyright infringement without restricting innovation. *Stanford Law Review* 56: 1345–1434.

Lichtman, D. 2000. Property rights in emerging platform technologies. *Journal of Legal Studies* 29: 615–648.

Lichtman, D. 2009. Copyright as innovation policy: Google Book Search from a law and economics perspective. In *Innovation policy and the economy*, vol. 9, ed. A.B.. Jaffe, J. Lerner, and S. Stern, 55–77. Chicago: University of Chicago Press.

Lichtman, D., and W.M. Landes. 2003. Indirect liability for copyright infringement: An economic perspective. *Harvard Journal of Law & Technology* 16: 395–410.

Liebowitz, S.J. 1985. Copying and indirect appropriability: Photocopying of journals. *Journal of Political Economy* 93: 945–957.

Liebowitz, S.J. 2006. Economists examine file sharing and music sales. In *Industrial organization and the digital economy*, ed. G. Illing and M. Peitz, 145–173. Cambridge: MIT Press.

Liebowitz, S.J. 2008. Testing file sharing's impact on music album sales in cities. *Management Science* 54: 852–859.

Martínez-Sánchez, F. 2010. Avoiding commercial piracy. *Information Economics and Policy* 22: 398–408.

Menell, P.S. 2009. Indirect copyright liability and technological innovation. *Columbia Journal of Law & the Arts* 32: 375–399.

Menell, P.S., and S. Scotchmer. 2007. Intellectual property law. In *Handbook of law and economics*, vol. 2, ed. A.M. Polinsky and S. Shavell), 1471–1570. Amsterdam: North-Holland.

Merges, R.P. 1997. The end of friction? Property rights and contract in the 'Newtonian' world of online-commerce. *Berkeley Technology Law Journal* 12: 115–136.

Michel, N.J. 2006. The impact of digital file sharing on the music industry: An empirical analysis. *Topics in Economic Analysis & Policy* 6: 18.

Müller-Langer, F., and M. Scheufen. 2011. The Google Book search settlement: A law and economics analysis. *Review of Economic Research on Copyright Issues* 8: 7–50.

Nandi, T.K., and F. Rochelandet. 2008. The incentives for contributing digital contents over P2P networks: An empirical investigation. *Review of Economic Research on Copyright Issues* 5: 19–35.

Netanel, N.W. 2003. Impose a noncommercial use levy to allow free peer-to-peer file sharing. *Harvard Journal of Law & Technology* 17: 1–84.

Novos, I.E., and M. Waldman. 1984. The effects of increased copyright protection: An analytic approach. *Journal of Political Economy* 92: 236–246.

Oberholzer-Gee, F., and K. Strumpf. 2007. The effect of file sharing on record sales: An empirical analysis. *Journal of Political Economy* 115: 1–42.
Oberholzer-Gee, F., and K. Strumpf. 2010. File sharing and copyright. In *Innovation policy and the economy*, vol. 10, ed. J. Lerner and S. Stern, 19–55. Chicago: University of Chicago Press.
Oliar, D. 2012. The copyright-innovation trade-off: Property rules, liabilty rules, and intentional infliction of harm. *Stanford Law Review* 64: 951–1020.
Park, Y., and S. Scotchmer. 2005. *Digital rights management and the pricing of digital products*, NBER working paper No. 11532. Cambridge, MA: National Bureau of Economic Research.
Peitz, M., and P. Waelbroeck. 2004. The effect of Internet piracy on music sales: Crosssection evidence. *Review of Economic Research on Copyright Issues* 1: 71–79.
Peitz, M., and P. Waelbroeck. 2006a. Piracy of digital products: A critical review of the theoretical literature. *Information Economics and Policy* 18: 449–476.
Peitz, M., and P. Waelbroeck. 2006b. Why the music industry may gain from free downloading: The role of sampling. *International Journal of Industrial Organization* 24: 907–913.
Rasch, A., and T. Wenzel. 2013. Piracy in a two-sided software market. *Journal of Economic Behavior & Organization* 88: 78–89.
Regner, T., and J.A. Barria. 2009. Do consumers pay voluntarily? The case of online music. *Journal of Economic Behavior & Organization* 71: 395–406.
Rob, R., and J. Waldfogel. 2006. Piracy on the high C's: Music downloading, sales displacement, and social welfare in a sample of college students. *Journal of Law & Economics* 49: 29–62.
Rob, R., and J. Waldfogel. 2007. Piracy on the silver screen. *Journal of Industrial Economics* 55: 379–395.
Rosso, M.A., and B.J. Jansen. 2010. Brand names as keywords in sponsored search advertising. *Communications of the Association for Information Systems* 27: 81–98.
Samuelson, P., and S. Scotchmer. 2002. The law and economics of reverse engineering. *Yale Law Journal* 111: 1575–1663.
Shy, O., and J.-F. Thisse. 1999. A strategic approach to software protection. *Journal of Economics & Management* 8: 163–190.
Smith, M.D., and R. Telang. 2009. Competing with free: The impact of movie broadcasts on DVD sales and Internet piracy. *MIS Quarterly* 33: 321–338.
Strahilevitz, L.J. 2003. Charismatic code, social norms, and the emergence of cooperation on thefile-swapping networks. *Virginia Law Review* 89: 505–595.
Sundararajan, A. 2004. Managing digital piracy: Pricing and protection. *Information Systems Research* 15: 287–308.
Takeyama, L.N. 1994. The welfare implications of unauthorized reproduction of intellectual property in the presence of demand network externalities. *Journal of Industrial Economics* 42: 155–166.
Takeyama, L.N. 2009. Copyright enforcement and product quality signaling in markets for computer software. *Information Economics and Policy* 21: 291–296.
Vernik, D.A., D. Purohit, and P.S. Desai. 2011. Music downloads and the flip side of digital rights management. *Marketing Science* 30: 1011–1027.
Waldfogel, J. 2010. Music file sharing and sales displacement in the iTunes era. *Information Economics and Policy* 22: 306–314.
Waldfogel, J. 2012a. Copyright protection, technological change, and the quality of new products: Evidence from recorded music since Napster. *Journal of Law & Economics* 55: 715–740.
Waldfogel, J. 2012b. Copyright research in the digital age. Moving from piracy to the supply of new products. *American Economic Review* 102: 337–342.
Waldfogel, J. 2012c. Digital piracy: Empirics. In *The oxford handbook of the digital economy*, ed. M. Peitz and J. Waldfogel), 512–546. Oxford: Oxford University Press.
Waldfogel, J. 2012d. Music piracy and its effects on demand, supply, and welfare. In *Innovation policy and the economy*, vol. 12, ed. J. Lerner and S. Stern, 92–109. Chicago: University of Chicago Press.
Wu, T. 2005. Copyright's communications policy. *Michigan Law Review* 103: 278–366.
Wu, S., and P. Chen. 2008. Versioning and piracy control for digital information goods. *Operations Research* 56: 157–172.
Yoon, K. 2002. The optimal level of copyright protection. *Information Economics and Policy* 14: 327–348.
Zentner, A. 2006. Measuring the effect of file sharing on music purchases. *Journal of Law & Economics* 49: 63–90.
Zentner, A. 2010. *Measuring the impact of file sharing on the movie industry: An empirical analysis using a panel of countries*. Manuscript available at http://ssrn.com/abstract=1792615. Accessed 10 May 2013.

Law of Demand

Michael Jerison and John K. -H. Quah

Abstract

We formulate several laws of individual and market demand and describe their relationship to neoclassical demand theory. The laws have implications for comparative statics and stability of competitive equilibrium. We survey results that offer interpretable sufficient conditions for the laws to hold and we refer to related

empirical evidence. The laws for market demand are more likely to be satisfied if commodities are more substitutable. Certain kinds of heterogeneity across individuals make the laws more likely to hold in the aggregate even if they are violated by individuals.

Keywords

Asymmetric information; Bernoulli utility function; Cobb–Douglas preferences; Comparative statics; Compensated demand; Engel curve; Giffen effects; Giffen goods; Income effect; Jacobian matrix; Law of demand; Lyapunov's second theorem; Marginal utility of income; Metonymy; Non-decreasing dispersion of excess demand; Portfolio choice; Risk aversion; Slutsky matrix; Stability of equilibrium; Substitution effect; Tâtonnement; Uniqueness of equilibrium

JEL Classifications

D1; D5; C62; D01; D11; D21; G11

The most familiar version of the law of demand says that as the price of a good increases the quantity demanded of the good falls. The principal use of the law of demand in economic theory is to provide sufficient and, in some contexts, necessary conditions for the uniqueness and stability of equilibrium, and for intuitive comparative statics. To guarantee such properties in equilibrium models with more than one good, the familiar one-good law of demand just stated is not sufficient – some multi-good version of the law is needed. In its multi-good form, the law of demand is said to hold for a particular change in prices if the prices and the quantities demanded move in opposite directions; in formal terms, the vector of price changes and the vector of resulting demand changes have a negative inner product.

In this article, we examine different formulations of the law of demand. They differ principally in the domain of price changes over which the law applies. It is not always the case that the law of demand is required to hold for *all* price changes: the version of the law which is required for stability analysis and comparative statics varies from one context to another. For each formulation of the law of demand, we discuss the conditions which are sufficient to guarantee that it is satisfied.

To point out the obvious, the law of demand, in whatever form, is not a universal law at all but a condition which may hold in some situations and not others. It is well known that, in transactions where asymmetric information is an important consideration, violations of the law can occur. For example, lowering the price of a set of used cars does not necessarily lead to higher demand if potential buyers think that the lower price reflects the quality of the cars being offered. (For a discussion of violations of the law of demand and other issues which arise when price has an impact on the perceived quality of the good being exchanged, see Stiglitz 1987.) In this article we make the classical assumption that the features of the good being transacted are commonly known and independent of the price. As we shall see, even in this classical setting various forms of the law of demand will hold only under conditions which are often neither obviously onerous nor obviously innocuous; in these cases, one must necessarily turn to empirical work to ascertain whether or not the law holds.

We use the notation and terminology of Mas-Colell et al. (1995, chs. 2, 3, 5) and assume that the reader is familiar with the basic consumer and producer theory described there. We assume that there are L commodities and that consumers are price-takers. The demand of a consumer of type α with income w at price vector $p = (p_\ell)_{\ell=1}^{L} \gg 0$ is the vector $x(p, w, \alpha) = (x_\ell(p, w, \alpha))_{\ell=1}^{L}$ in $\mathbb{R}_+^L$, satisfying the budget identity $p \cdot x(p, w, \alpha) = w$ for all p and w. Unless stated otherwise, we assume the demand function $x(\cdot, \cdot, \alpha)$ to be C^1. Then it has a Slutsky matrix of substitution effects $S(p,w,\alpha)$ with ℓj element $S_{\ell j}(p, w, \alpha) = \partial x_\ell(p, w, \alpha)/\partial p_j + [\partial x_\ell(p, w, \alpha)/\partial w]x_j(p, w, \alpha)$ The Slutsky matrix $S(p,w,\alpha)$ is the Jacobian matrix of the Slutsky-compensated demand function x^*, defined by $x^*(q) = x(q, q \cdot x(p, w, \alpha), \alpha)$, evaluated at $q = p$. The term $[\partial x_\ell(p, w, \alpha)/\partial w]x_j(p, w, \alpha)$ is called an income effect since it approximates the effect on the demand for good ℓ when income rises enough to compensate for a unit increase in the price of good j. If the consumer chooses demand

bundles by maximizing a well-behaved utility function, then the Slutsky matrix is symmetric and negative semidefinite. The latter means that $v \cdot S(p, w, \alpha)v \leq 0$ *for all* $v \in \mathbb{R}^L$; in particular, the diagonal terms of the Slutsky matrix are non-positive.

One-Good and Multi-good Laws of Demand

The term 'law of demand' most often refers to the effect of price changes on consumers with fixed incomes. The law for a single good ℓ and a single consumer of type α is

$$(p_\ell - \overline{p}_\ell)(x_\ell(p, w, \alpha) - x_\ell(\overline{p}, w, \alpha)) \leq 0, \qquad (1)$$

for p and $\overline{p}$, with $p_i = \overline{p}_i$ for all $i \neq \ell$ and income w fixed. (In the *strict* version of the law, the weak inequality in (1) is replaced by strict inequality when $p \neq \overline{p}$; all the laws of demand discussed in this article can be stated in their corresponding strict forms, though we generally do not do so.) The inequality (1) is equivalent to

$$\begin{aligned} 0 \geq & \frac{\partial x_\ell}{\partial p_\ell}(p, w, \alpha) \\ & = S_{\ell\ell}(p, w, \alpha) \\ & \quad - x_\ell(p, w, \alpha)\frac{\partial x_\ell}{\partial w}(p, w, \alpha), \forall (p, w). \end{aligned}$$

It holds if the substitution effect $S_{\ell\ell}$ is negative and larger in magnitude than the income effect $x_\ell(p, w, \alpha)\frac{\partial x_\ell}{\partial w}(p, w, \alpha)$. If the consumer is utility-maximizing, then $S_{\ell\ell} \leq 0$, so a sufficient condition for good ℓ to obey the law of demand is that the demand for this good is normal ($\partial x_\ell(p, w, \alpha)/\partial w \geq 0$). If the demand for good ℓ is not normal, the price effect $\partial x_\ell/\partial p_\ell$ may be positive. This is called a Giffen effect and good ℓ is called a Giffen good. All goods are normal and Giffen effects are ruled out if the demand function is generated by homothetic preferences or by a concave additive utility function $\left(u(x) = \sum_{\ell=1}^{L} u_\ell(x_\ell)\right)$, or, more generally, by a supermodular concave function u, that is, one in which all commodity pairs are Auspitz–Lieben–Edgeworth–Pareto complements: $\partial^2 u(x)/\partial x_j \partial x_\ell \geq 0$ for all $j \neq \ell$ (Chipman 1977).

Giffen goods are rarely observed. Sometimes demand for a durable good like oil may increase with its current price if traders expect an even higher price in the future. However, if commodities are distinguished by date, this is not a Giffen effect since a future price changes along with the current price. A possible example of a Giffen good is proposed by Baruch and Kannai (2002). They give evidence suggesting that, in Japan of the 1970s, shochu, a cheap (and, by some accounts, nasty) alcoholic drink, fits the definition. One may explain the demand for shochu in the following way. A consumer chooses between sake (good 1) and shochu (good 2). He always prefers sake to shochu, but he also *must have* a minimum alcohol intake (which we fix at 1). Formally, his utility is $u(x_1, x_2) = x_1$, subject to the 'survival' constraint $x_1 + x_2 \geq 1$. If the consumer is sufficiently poor, both the budget and survival constraints bind, with the consumer consuming as much sake – and as little shochu – as possible. A fall in the price of shochu allows him to buy less shochu and more sake and still meet his alcohol requirement; this he chooses to do since he always prefers sake to shochu.

Turning now to multi-good laws of demand, let $P \subseteq \mathbb{R}^L_{++}$ be a set of prices and let $X : P \to \mathbb{R}^L$ be a function representing individual or aggregate demand of firms or of consumers. The natural multi-good generalization of the one-good law in (1) is

$$(p - p') \cdot (X(p) - X(p')) \leq 0 \qquad (2)$$

for all (p, p') in some subset of $P \times P$. If P is convex and open and X is C^1, (2) holds on $P \times P$ if and only if the Jacobian matrix $\partial X(p)$ is negative semidefinite at each p (Hildenbrand and Kirman 1988).

Suppose that the supply vector of the L goods changes from ω to ω'. Let p and p' be corresponding equilibrium prices so $X(p) = \omega$ and $X(p') = \omega'$. Then, if X obeys (2) for all prices, we obtain $(p - p') \cdot (\omega - \omega') \leq 0$. It is clear that this comparative statics property and the law of

demand on X are essentially two sides of the same coin. Note also that, according to this property, an increase in the supply of good k, with the supply of all other goods held fixed, will lead to a fall in the price of k.

Suppose that P is open and X obeys the *strict* law of demand, that is, X satisfies (2) with strict inequality for all distinct p and p' in P. This implies in particular that X is $1-1$ and that, for each $\overline{\omega}$ in $X(P)$, there is a unique equilibrium price vector $\overline{p} = X^{-1}(\overline{\omega})$. A tâtonnement path for the function $X - \overline{\omega}$ is the solution to $dp/dt = X(p(t)) - \overline{\omega}$ for some initial condition $p(0) = p'$ in P. We say that $X - \overline{\omega}$ is monotonically stable for $\overline{\omega}$ if each of its tâtonnement paths satisfies $d|p(t) - \overline{p}|^2/dt < 0$ whenever $p(t) \neq \overline{p}$. It is easy to check that $X - \overline{\omega}$ is monotonically stable for all $\overline{\omega}$ in $X(P)$ if and only if X obeys the strict law of demand. Furthermore, because P is open, a tâtonnement path for $X - \overline{\omega}$ which begins at a price sufficiently close to $\overline{p} = X^{-1}(\overline{\omega})$ stays in P for all $t > 0$. Lyapunov's second theorem then guarantees that the tâtonnement path converges to $\overline{p}$.

Laws of demand are thus useful as intuitive sufficient conditions for the uniqueness and stability of equilibrium and for comparative statics. We will examine, in different contexts, circumstances under which they hold.

Law of Demand for Competitive Firms and Consumers with Quasilinear Utility

For a firm with production set Y, profit maximizing net output vector y at price vector p and $\overline{y}$ at $\overline{p}$ satisfy $p \cdot y \geq p \cdot \overline{y}$ and $\overline{p} \cdot \overline{y} \geq \overline{p} \cdot y$. The net demand vectors $x = -y$ and $\overline{x} = -\overline{y}$ satisfy $p(x - \overline{x}) \leq 0$ and $\overline{p}(x - \overline{x}) \geq 0$, hence satisfy the law of demand: $(p - \overline{p}) \cdot (x - \overline{x}) \leq 0$. Similarly, a consumer with utility function $u(x_0, x) = x_0 + \varphi(x_1, \ldots, x_L)$ (*quasilinear* with respect to good 0) and with sufficiently high income w satisfies the law of demand on a restricted domain, where the price of good 0 is fixed (say at 1). This is a special case of the law for firms. The consumer's optimal demand for goods 1 through L at p (the price vector for goods 1 to L) and income w maximizes $w - p \cdot x + \varphi(x)$. This is equivalent to profit maximization with x an input vector and $\varphi(x)$ the value of output.

Bewley (1977) shows that a long-lived consumer with a random income stream and a random but stationary time-separable utility function, who is constrained from borrowing, will accumulate savings so that the marginal utility of income is nearly constant. In the short run, this consumer acts (nearly) as if its utility is quasilinear with respect to money, and its short run demands for other goods satisfy the law of demand. Vives (1987) formalizes Marshall's idea (in his *Principles*) that consumer demands for goods with small expenditure shares are close to demands generated by quasilinear utility.

Multi-Good Laws of Demand for a Consumer

Suppose the demand of a consumer of type α is determined by maximizing a utility function u^α. The Hicksian compensated demand $h(p, \bar{u}, \alpha)$ is a bundle that minimizes $p \cdot x$ subject to $u^\alpha(x) \geq \bar{u}$. Keeping the utility level fixed at $\bar{u}$, this Hicksian demand function satisfies the multi-good law of demand: (2) holds for $X(p) = h(p, \bar{u}, \alpha)$. Utility maximization also guarantees that $x(\cdot, \cdot, \alpha)$ satisfies the *weak weak axiom of revealed preference*: $p \cdot x(p', w', \alpha) \leq w \Rightarrow p' \cdot x(p, w, \alpha) \geq w'$. Equivalently, for any fixed w, $X(p) = x(p, w, \alpha)$ satisfies (2) on the restricted domain with $p \cdot X(p') = w$. This is also called the compensated law of demand since the demand vector $X(p')$ remains barely affordable when the price vector changes from p' to p. The weak weak axiom is satisfied so long as the consumer maximizes a *complete* preference relation; the preferences need not be transitive. When $x(\cdot, \cdot, \alpha)$ is C^1, the following are equivalent: (i) $x(\cdot, \cdot, \alpha)$ obeys the weak weak axiom; (ii) its Slutsky matrix $S(p,w,\alpha)$ is negative semidefinite (but not necessarily symmetric); (iii) its Jacobian matrix $\partial_p x(p,w,\alpha)$ is negative semidefinite on the hyperplane orthogonal to $x(p,w,\alpha)$ (Kihlstrom et al. 1976; Brighi 2004).

When we say that $x(\cdot, \cdot, \alpha)$ obeys the unrestricted law of demand (or law of demand, for short) we mean that for each w, $X(p) = x(p, w,$

α) satisfies (2) for *all* price changes. Since this is equivalent to negative semidefiniteness of the Jacobian $\partial_p x(p,w,\alpha)$ for all p, it is stronger than simply saying that the diagonal terms of the matrix are non-positive. Thus it is not equivalent to the one-good law of demand for every good and does not follow from the assumption that the demand for every good is normal.

Let $M(p, w, \alpha)$ be the income effects matrix, with ℓj component $[\partial_w x_\ell(p, w, \alpha)]x_j(p, w, \alpha)$. From the Slutsky decomposition, $\partial_p x(p, w, \alpha) = S(p, w, \alpha) - M(p, w, \alpha)$, we see that type α satisfies the law of demand if it satisfies the weak weak axiom and $M(p, w, \alpha)$ is positive semidefinite at each p. However, the latter condition is strong; it occurs if and only if demand is linear in income for all goods, which excludes the possibility of luxuries or necessities.

A more promising approach is to find conditions under which the Slutsky matrix always 'dominates' the income effects matrix even when the latter 'misbehaves'. On the assumption that type α has a concave utility function u^α, a sufficient and (in a sense) necessary condition for the law of demand is $-[x^T\partial^2 u^\alpha(x)x]/(\partial u^\alpha(x)x) \leq 4, \forall x$. This result was obtained independently by Milleron (1974) and Mitjuschin and Polterovich (1978) (see also Mas-Colell et al. 1995, p. 145, and an alternative formulation in Kannai 1989).

An important application of this result is in the theory of portfolio choice. In that case, the demand bundle is the consumer's contingent consumption over L states of the world; it is standard to assume that the consumer has a von Neumann–Morgenstern utility function $u^\alpha(x) = \sum_{i=1}^{l} \pi_i v^\alpha(x_i)$, where π_i is the subjective probability of state i and $v^\alpha{:}R_{+\,+} \to R$ is the Bernoulli utility function. Suppose the coefficient of relative risk aversion, $-yv^{\alpha''}(y)/v^{\alpha'}(y)$, does not *vary* by more than four on the domain of v^α. Then the consumer's demand for contingent consumption at different state prices will obey the law of demand; this in turn implies that the law of demand holds for the consumer's demand for securities, whether or not the market is complete (Quah 2003).

Laws of Market Demand When the Income Distribution Is Independent of Price

Consider a large economy with consumers drawn at random from a probability space $A \times \mathbb{R}_+$ of consumer types and their incomes, with distribution μ. The expected aggregate (market) demand vector at prices p is $X(p) = \int_{A\times\mathbb{R}_+} x(p, w, \alpha)d\mu$. We are interested in conditions under which X obeys the unrestricted law of demand, that is, (2) holds for all price changes; equivalently, $\partial X(p)$ is negative semidefinite for all p. If $x(\cdot, \cdot, \alpha)$ obeys the law of demand for all α, then, clearly, so will X. One justification for studying the law of demand at the individual level is that it is preserved by aggregation.

Aggregating the Slutsky decomposition across all agents, the law of demand requires

$$v \cdot \partial X(p)v = v \cdot \left[\int_{\alpha\in A} x(p, w, \alpha)d\mu\right]v = v \cdot \overline{S}(p)v - v \cdot \overline{M}(p)v \leq 0, \forall v \quad (3)$$

where $\overline{S}(p) = \int S(p, w, \alpha)d\mu$ is the mean Slutsky matrix, and $\overline{M}(p)$ is the mean income effects matrix, with ℓj element $\int[\partial x_\ell(p, w, \alpha)/\partial w]x_j(p, w, \alpha)d\mu$. (We assume here and below that these integrals exist.) If all consumers obey the weak weak axiom, which they do if they are utility maximizers, then $S(p,w,\alpha)$ and hence $\overline{S}(p)$ are negative semidefinite; so $\partial X(p)$ is negative semidefinite if $\overline{M}(p)$ is positive semidefinite.

The matrix $\overline{M}(p)$ is determined by the consumers' Engel curves $x(p \cdot, \cdot, \alpha)$ at p. Positive semidefiniteness of this matrix is known as *increasing spread* (Hildenbrand 1994). To see why, note that

$$2v \cdot \overline{M}(p)v = \partial_t \int [v \cdot x(p, w + t, \alpha)]^2 d\mu(\alpha, w)|_{t=0}. \quad (4)$$

We can interpret $v \cdot x(p,w,\alpha)$ as α's demand for a commodity (call it T_v), which is consumed when the other goods are consumed; specifically, the

L

consumption of one unit of good j requires v_j units of T_v. Then $\int[v \cdot x(p, w, \alpha)]^2 d\mu$ measures the spread of the consumers' demands for T_v around the origin. By (4), $\overline{M}(p)$ is positive semidefinite if and only if for every v the consumers' demands for T_v spread out from 0 as their incomes rise. This is the multi-good generalization of normality, where the consumers' demands for a single good increase (spread from 0) as their incomes rise.

We now consider various interpretable conditions on the distribution of consumer characteristics which guarantee increasing spread (and thus the law of demand). This property holds if consumers have the same demand function and income is distributed with a non-increasing density function ρ on $[0,\overline{w}]$ (Hildenbrand 1983). In that case, integrating by parts, (4) becomes $2v \cdot \overline{M}(p)v = [v \cdot x(p, \overline{w}, \alpha)]^2 \rho(\overline{w}) - \int[v \cdot x(p, w, \alpha)]^2 \rho'(w)dw \geq 0$. While the non-increasing density condition is strong, imposing some weak restrictions on the Engel curves will guarantee increasing spread for a significantly larger class of income density functions (Chiappori 1985). However, to guarantee increasing spread for *every* non-trivial income distribution requires stringent conditions on the consumers' Engel curves: $x(p, \cdot, \alpha)$ must lie in a single plane (depending on p) and the demand for each good is either a concave or convex function of income (Freixas and Mas-Colell 1987; Jerison 1999).

Increasing spread is also implied by certain kinds of behavioural heterogeneity across consumers. We consider consumers with the same income w and demands of the form $x_\ell(p, w, \alpha) = e^{\alpha_\ell}\hat{x}(e^{\alpha_1}p_1, \ldots, e^{\alpha_L}p_L, w)$, where $\hat{x}$ is an arbitrary demand function and $\alpha = (\alpha_1, \ldots, \alpha_L) \in R^L$. If $\hat{x}$ is generated by some utility function $\hat{u}$, then $x(\cdot, \cdot, \alpha)$ is generated by the utility function $u^\alpha(x) = \hat{u}(e^{-\alpha_1}x_1, \ldots, e^{-\alpha_L}x_L)$. Increasing spread is guaranteed if α has a sufficiently flat density over R^L. This condition also ensures that the mean Slutsky matrix $\overline{S}(p)$ is negative semidefinite even if $\hat{x}$, hence each $x(\cdot, \cdot, \alpha)$, violates the weak weak axiom (and so is *not* generated by a utility function). Thus when α has a sufficiently flat density, X satisfies the law of demand; in fact it can be shown that X is nearly generated by Cobb–Douglas preferences (Grandmont 1992). Whether flatness of the α density implies heterogeneity (in some meaningful sense) of the consumers' demands depends on the behaviour of $\hat{x}$ (Giraud and Quah 2003).

Even when $\overline{M}(p)$ is not positive semidefinite, that is, $v \cdot \overline{M}(p)v < 0$ for some v, it is clear from (3) that $v \cdot \partial X(p)v < 0$ can hold provided the substitution effects are large enough, that is, $v \cdot \overline{S}(p)v$ is sufficiently negative. This feature can be exploited; for example, one can substantially weaken the non-increasing density condition in Hildenbrand (1983; described above) and still obtain the law of demand if substitution effects are accounted for through restrictions on the utility function (Quah 2000). Similarly, a large enough positive income effect can compensate for consumers' violations of the weak weak axiom, that is, situations where, for some v, $v \cdot \overline{S}(p)v > 0$.

Whether the substitution effect $v \cdot \overline{S}(p)v$ dominates the income effect $v \cdot \overline{M}(p)v$ is an empirical question. The sizes of the effects must be estimated. Härdle et al. (1991) show how this can be done with cross-section data under standard econometric assumptions, without restrictions on the functional forms of the consumer demands. In most empirical demand analyses, consumers are grouped according to observable attributes other than income, and within a group, a, the consumers' budget share vectors are assumed to have the form $b^a(p, w) + \varepsilon$, where ε is a mean 0 random variable with distribution independent of income w. Under this assumption, a consumer's type is its attribute group and a realized value of ε. Within group a, the distribution of types with income w, denoted $\mu^a(\alpha|w)$, does not vary with w. Thus, if the income distribution in the group has a density ρ^a, then

$$\int \left\{\partial_w[v \cdot x(p, w, \alpha)]^2\right\} d\mu^a = \int \left\{\partial_w \int [v \cdot x(p, w, \alpha)]^2 d\mu^a(\alpha|w)\right\} \rho^a(w)dw, \forall v \in R^L. \tag{5}$$

The left side of (5) equals $2v \cdot M^a(p)v$, where $M^a(p)$ is the mean income effect matrix of the consumers in group a. The right side of (5) is the

mean of the derivative of $\int [v \cdot x(p, w, \alpha)]^2 d\mu^a(\alpha|w)$ with respect to w. It can be efficiently estimated by the nonparametric method of average derivatives (Härdle and Stoker 1989). The mean income effect matrix $\overline{M}(p)$ is a weighted average of the matrices $M^a(p)$, weighted by the shares of the population in the groups a. Condition (5), called metonymy, is weaker than the assumption that the budget shares have the form $b^a(p, w) + \varepsilon$, so weak, in fact, that it is not potentially refutable with infinite cross-section data (Evstigneev et al. 1997; Jerison 2001). Income effect matrices estimated in this way using cross-section expenditure data from several countries are all positive semi-definite (Härdle, Hildenbrand and Jerison 1991; Hildenbrand and Kneip 1993).

Laws of Demand in Private Ownership Economies

In the previous section, we assumed consumer incomes to be exogenously given independently of prices. This is plainly not true in general equilibrium. For example, consider a private ownership economy with consumers drawn randomly from a distribution μ over types, where type α has the demand function $x(\cdot, \cdot, \alpha)$ and an endowment vector ω^α. If the consumers receive no profits, the income of type α at price vector p is $p \cdot \omega^\alpha$. We are interested in laws of demand that can be satisfied by the consumer sector's aggregate demand $\tilde{X}(p) = \int x(p, p \cdot \omega^\alpha, \alpha)\, d\mu$ or aggregate excess demand $\zeta(p) = \tilde{X}(p) - \overline{\omega}$, where $\overline{\omega} = \int \omega^\alpha d\mu$ is the aggregate endowment.

The first thing to note is that under standard assumptions, both $\tilde{X}$ and ζ are zero-homogeneous and, essentially for this reason, satisfy the unrestricted law of demand only in exceptional cases (Hildenbrand and Kirman 1988). However, if the consumers' endowments are collinear (that is, if for each α there is some $k \geq 0$ with $\omega^\alpha = k\overline{\omega}$) then the sufficient conditions for the law of market demand given in the previous section are also sufficient for $\tilde{X}$ (and hence ζ) to satisfy (2) for p and p' in $P = \{p \in R^L_{++} : p \cdot \overline{\omega} = 1\}$; in other words, the law of demand holds for mean income preserving price changes. This is so because, when endowments are collinear, a price change which preserves mean income also preserves the income of *every* agent.

When we drop the strong assumption of collinear endowments, this restricted form of the law of demand is not guaranteed even if all consumers have homothetic preferences (Mas-Colell et al. 1995, p. 598). However, it *does* hold when the consumer sector has two properties: (*a*) all agents have homothetic preferences and (*b*) the preferences and endowments are independently distributed. Quah (1997) shows that this scenario can be understood as the idealization of a more general situation. The crucial feature of homothetic preferences here is that they generate demand functions which are linear in income. Retaining the independence assumption (*b*), one can show that, when substitution effects are non-trivial (in some specific sense), $\tilde{X}$ obeys the restricted law of demand provided the mean demand of agents with identical endowments is not 'too non-linear' in income. This last property can arise from an appropriate form of heterogeneity in demand behaviour, which can be modelled using the parametric framework employed by Grandmont (1987, 1992).

It is interesting to ask when aggregate consumer excess demand ζ satisfies the weak weak axiom: $p \cdot \zeta(p') \leq 0 \Rightarrow p' \cdot \zeta(p) \geq 0$. This condition ensures that the set of equilibrium prices is convex in all competitive production economies with convex technology and constant returns to scale; furthermore, it is the weakest restriction on ζ guaranteeing this conclusion (Mas-Colell et al. 1995, p. 609). The sufficiency of this condition hinges on the fact that the production side of the economy satisfies the law of demand. Since the equilibrium set is generically discrete, its convexity implies generic uniqueness of equilibrium (up to scalar multiple). When ζ satisfies the weak weak axiom it also satisfies the law of demand (2) on the restricted set with $p \cdot \zeta(p') = 0$. If (2) holds strictly on this set when p and p' are not collinear, then the unique equilibrium is globally stable under tâtonnement, and there are natural comparative statics.

With the use of a Slutsky decomposition, it can be shown that ζ satisfies the weak weak axiom if

the mean Slutsky matrix $S(p)$ is negative semi-definite (as it is if the consumers are utility maximizing) and the consumers' excess demand vectors spread apart on average when their incomes rise. The latter condition is called non-decreasing dispersion of excess demand (NDED). To formalize it, define $z(p, t, \alpha) \equiv x(p, t + p \cdot \omega^{\alpha}) - \omega^{\alpha}$, the excess demand of type α with income transfer t. The corresponding aggregate excess demand is $Z(p, t) \equiv \int z(p, t, \alpha) d\mu$. NDED holds if $\partial_t \int \{v \cdot [z(p, t, \alpha) - Z(p, t)]\}^2 d\mu|_{t=0} \geq 0$ for every $p \in R^L_{++}$ and every v with $v \cdot p = 0$ and $v \cdot \zeta(p) = 0$; in other words, the income transfers raise the variance of the composite excess demands $v \cdot z(p, t, \alpha)$ (Jerison 1999). Quah's 1997 model (described above) is an example of an economy where NDED is satisfied approximately.

See Also

- Comparative Statics
- Engel Curve
- General Equilibrium
- Giffen's Paradox
- Revealed Preference Theory
- Risk Aversion
- Tâtonnement and Recontracting

Bibliography

Baruch, S., and Y. Kannai. 2002. Inferior goods, Giffen goods, and shochu. In *Economics essays*, ed. G. Debreu, W. Neuefeind, and W. Trockel. Berlin: Springer.

Bewley, T. 1977. The permanent income hypothesis: A theoretical formulation. *Journal of Economic Theory* 16: 252–292.

Brighi, L. 2004. A stronger criterion for the weak weak axiom. *Journal of Mathematical Economics* 40: 93–103.

Chiappori, P.-A. 1985. Distribution of income and the 'law of demand'. *Econometrica* 53: 109–128.

Chipman, J. 1977. An empirical implication of Auspitz–Lieben–Edgeworth–Pareto complementarity. *Journal of Economic Theory* 14: 228–231.

Evstigneev, I.V., W. Hildenbrand, and M. Jerison. 1997. Metonymy and cross-section demand. *Journal of Mathematical Economics* 28: 397–414.

Freixas, X., and A. Mas-Colell. 1987. Engel curves leading to the weak axiom in the aggregate. *Econometrica* 55: 515–531.

Giraud, G., and J.K.H. Quah. 2003. Homothetic or Cobb–Douglas behavior through aggregation. *Contributions to Theoretical Economics* 3(1), Article 8.

Grandmont, J.M. 1987. Distribution of preferences and the 'law of demand'. *Econometrica* 55: 155–161.

Grandmont, J.M. 1992. Transformations of the commodity space, behavioral heterogeneity, and the aggregation problem. *Journal of Economic Theory* 57: 1–35.

Härdle, W., W. Hildenbrand, and M. Jerison. 1991. Empirical evidence on the law of demand. *Econometrica* 59: 1525–1549.

Härdle, W., and T. Stoker. 1989. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84: 986–995.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51: 997–1019.

Hildenbrand, W. 1994. *Market demand*. Princeton: Princeton University Press.

Hildenbrand, W., and A. Kirman. 1988. *Equilibrium analysis*. Amsterdam: North-Holland.

Hildenbrand, W., and A. Kneip. 1993. Family expenditure data, heterscedasticity and the law of demand. *Ricerche Economiche* 47: 137–165.

Jerison, M. 1999. Dispersed excess demands, the weak axiom, and uniqueness of equilibrium. *Journal of Mathematical Economics* 31: 15–48.

Jerison, M. 2001. Demand dispersion, metonymy and ideal panel data. In *Economics essays*, ed. G. Debreu, W. Neuefeind, and W. Trockel. Berlin: Springer.

Kannai, Y. 1989. A Characterization of monotone individual demand functions. *Journal of Mathematical Economics* 18: 87–94.

Kihlstrom, R., A. Mas-Colell, and H. Sonnenschein. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44: 971–978.

Mas-Colell, A. 1991. On the uniqueness of equilibrium once again. In *Equilibrium theory and applications*, ed. W. Barnett et al. Cambridge: Cambridge University Press.

Mas-Colell, A., M.D. Whinston, and J.R. Green. 1995. *Microeconomic theory*. Oxford: Oxford University Press.

Milleron, J.C. 1974. Unicité et stabilité de l'équilibre en économie de distribution. Unpublished seminar paper, Séminaire d'Econométrie Roy-Malinvaud. Paris: CNRS.

Mitjuschin, L.G., and W.M. Polterovich. 1978. Criteria for monotonicity of demand functions. *Ekonomika i Matematicheskie Metody* 14: 122–128.

Quah, J.K.-H. 1997. The law of demand when income is price dependent. *Econometrica* 65: 1421–1442.

Quah, J.K.-H. 2000. The monotonicity of individual and market demand. *Econometrica* 68: 911–930.

Quah, J.K.-H. 2003. The law of demand and risk aversion. *Econometrica* 71: 713–721.

Stiglitz, J.E. 1987. The causes and consequences of the dependence of quality on price. *Journal of Economic Literature* 25(1): 1–48.

Vives, X. 1987. Small income effects: A Marshallian theory of consumer surplus and downward sloping demand. *Review of Economic Studies* 54: 87–103.

Law of Indifference

F. Y. Edgeworth

Abstract

A designation applied by Jevons to the following fundamental proposition: 'In the same open market, at any one moment, there cannot be two prices for the same kind of article.'

Keywords

Law of indifference

JEL Classifications

D0

A designation applied by Jevons to the following fundamental proposition: 'In the same open market, at any one moment, there cannot be two prices for the same kind of article.'

This proposition, which is at the foundation of a large part of economic science, itself rests on certain ulterior grounds: namely, certain conditions of a perfect market. One is that monopolies should not exist, or at least should not exert that power in virtue of which a proprietor of a theatre, in Germany for instance, can make a different charge for the admission of soldiers and civilians, of men and women. The indivisibility of the articles dealt in appears to be another circumstance which may counteract the law of indifference in some kinds of market, where price is not regulated by cost of production.

[Jevons (1875), *Theory of Exchange*, 2nd edn, p. 99 (statement of the law). Walker (1886), *Political Economy*, art. 132 (a restatement). Mill (1848), *Political Economy*, bk. ii. ch. iv. § 3 (imperfections of actual markets). Edgeworth (1881), *Mathematical Psychics*, pp. 19, 46 (possible exceptions to the law of indifference).]

Bibliography

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Jevons, W.S. 1875. *Money and the mechanism of exchange*. London: C. Kegan Paul & Co..

Mill, J.S. 1848. *Principles of political economy*. London: J.W. Parker.

Walker, F.A. 1886. *A brief textbook of political economy*. London.

Law(s) of Large Numbers

Werner Ploberger

Abstract

It is a well-known fact that averages of most random variables converge. The laws of large numbers are mathematical theorems which explain this phenomenon. We discuss the various forms of this theorem. Generalizations to dependent variables (ergodic ths) are introduced. We also mention uniform laws of large numbers, which are quite indispensable tools to prove consistency of estimators.

Keywords

Bernoulli experiments; Bernoulli, J.; Ergodic theorems; Law of large numbers; Maximum likelihood; Poisson, S. D.; Probability; Strong law of large numbers; Variance; Weak law of large numbers

JEL Classifications

C10

When we have a large number of independent replications of a random experiment, we observe that the frequency of the outcomes can be very well approximated by the probabilities of the corresponding events. The profits of many commercially successful enterprises – like casinos or insurance companies – are based on random events obeying some laws.

Mathematically, this idea was first formulated by Jacob Bernoulli, for experiments with only two outcomes ('Bernoulli experiments'). The terminology 'law of large numbers' was introduced by S.D. Poisson in 1835.

In the most basic version, LLN (the standard abbreviation for 'law(s) of large numbers') describes results of the following type. We assume that we have given a sequence of random variables X_1, X_2,... We say we have a LLN if

$$\frac{1}{N}(X_1 + ...X_N) \tag{1}$$

converges for $N \to \infty$, preferably to a constant.

For stating our results, we have to state the nature of the convergence in our LLN and impose some restrictions on the X_i. The more we restrict our X_i, the stronger our convergence results will be.

The Weak Law of Large Numbers

The 'weak law of large numbers' states that averages like (1) converge in a 'weak' sense (like for example convergence in probability) to a limit. In most cases, the requirements for the random variables involved are not very restrictive. A typical weak LLN is the following theorem.

Theorem 1 *Assume that the random variables X_i satisfy*

$$EX_i = 0, \tag{2}$$

$$\sup EX_i^2 < \infty \tag{3}$$

and

$$\lim_{M\to\infty} \sup_{|i-j|} > M \, |EX_iX_j| < \infty. \tag{4}$$

Then for $N \to \infty$

$$\frac{1}{N}(X_1 + ...X_N) \to^P 0,$$

where $\to^P$ denotes convergence in probability.

Our random variables have to be centred, of bounded variance, and condition (4) requires that the correlation of random variables 'far apart' converges to zero uniformly. This is a very general and important result. Another advantage is the simplicity of its proof: it is an elementary task to show that the variance of the average converges to zero. Then the theorem is an immediate consequence of Chebyshev's inequality. Moreover, the assumptions of the theorem can easily be checked, and only depend on the second moments of the X_i.

The Strong Law of Large Numbers

In some cases, we want to have more than convergence in probability of the averages. For this purpose, we have strong laws of large numbers. We do need, however, stricter requirements. The following theorem is a typical strong LLN. A more stringent discussion of this type of theorems

Theorem 2 *Assume that the random variables X_i satisfy* (2),(3). *Let $\mathfrak{F}_i$ be an increasing sequence of σ-algebras (for example $\mathfrak{F}_{i-1} \subset \mathfrak{F}_i$) so that X_i is $\mathfrak{F}_i$ –measurable. Then let us assume that*

$$E(X_i/\mathfrak{F}_{i-1}) = 0. \tag{5}$$

Then

$$\frac{1}{N}(X_1 + ...X_N) \to 0 \; P - almostsurely.$$

Heuristically, we can interpret $\mathfrak{F}_i$ as information available at time i. Then (5) postulates that we cannot predict X_i given the information at time $i - 1$. One important special case where (5) is fulfilled is the case of independent. In this case, we can choose $\mathfrak{F}_i$ to be the σ-algebra generated by X_1,... X_i,. Then, assuming the X_i to be independent, we have $E(X_i{=}\mathfrak{F}_{i-1}){=} E(X_i)$:

Hence (5) is more general than the requirement of independence, but still far more restrictive than (4).

Ergodic Theorems

We can easily see that (5) implies that our X_i are uncorrelated. In many applications, this requirement is unrealistic. Fortunately, there is a theory

guaranteeing convergence of sums like (1) at least for stationary processes X_i. A process X_i, $i \in \mathbf{Z}$ is called (strictly) stationary if for all $n \in \mathbf{Z}$ the distributions of $(X_1, X_2,... X_m)$ and $(X_{n+1}, X_2,...X_{n+m})$ are the same. To describe the limits of our process, we need to introduce the transition operator T: This operator is a mapping defined on the space of random variables measurable with respect to the σ-algebra generated by the X_i, $i \in \mathbf{Z}$. For random variables

$$Y = f(X_{t_1}, X_{t_2}, ...X_{t_n}) \tag{6}$$

we define the random variable TY by

$$TY = f(X_{t_1+1}, X_{t_2+1}, ...X_{t_n+1}). \tag{7}$$

So the transition operator T shifts every random variable 'one step in the future'. (T can be considered as the inverse of the usual lag operator). One can show that the definition based on (6), (7) can be uniquely extended to the space of all X_i, $i \in$ Z measurable random variables. Then an event A is called *invariant* if

$$TI_A = I_A \text{ almost surely,}$$

where I_A is the indicator of the event A. It can be easily seen that the invariant events form a $\sigma-$algebra, which we denote by $\mathfrak{F}$. Then the ergodic theorem states that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = E(X_i/\mathfrak{F}). \tag{8}$$

(Since we are taking the conditional expectation with respect to $\mathfrak{F}$, it can easily be seen that $E(X_1/\mathfrak{F}) = E(X_2/\mathfrak{F}) = ...$).

The ergodic theorem is included in most of advanced textbooks on probability theory (see, for example, Billingsley 1995). A more detailed exposition can be found in Gray (2007).

We now can take various conclusions from our theorem. First of all, we can regardless of the nature of the $\sigma-$algebra $\mathfrak{F}$ conclude that the limit of $\frac{1}{n}\sum_{i=1}^{n} X_i$ *exists*. In econometric theory, one often postulates the existence of limits of certain averages (that is, in regression theory we often assume that $\lim_{n\to\infty} n \sum_{i=1}^{n} x_i x_i'$ exists). In case of stationary processes, the theorem here makes assumptions of this type very plausible.

If the $\sigma-$algebra $\mathfrak{F}$ is trivial (that is, consists only of events of probability 0 and 1), then the right-hand side of (8) is constant. One sufficient criterion for this property is that the process is a causal function of i.i.d. random variables. So if

$$X_i = f(e_i, e_{i-1},...)$$

where e_i are i.i.d., $\mathfrak{F}$ is trivial.

Applications and Uniform Laws of Large Numbers

For many statistical applications, we need stronger results. As a first example, consider the asymptotic of the maximum likelihood estimator. As a simplest case, let us discuss the case of i.i.d. random variables X_i, distributed according to densities f_θ for parameters $\theta \in \Theta$, and let θ_0 be the true parameter. Then the LLN guarantees that for every fixed θ

$$\frac{1}{n} \sum \ln(f_\theta(X_i) \to \int \ln(f_\theta) f_{\theta_0}, \tag{9}$$

and the function on the right-hand side is maximized if $\theta = \theta_0$. Since the maximum likelihood estimator maximizes the right-hand side, it seems reasonable to exploit this relation for a proof of consistency of the maximum likelihood estimator. The LLN guarantees only convergence for *fixed* θ, from our LLN we cannot say anything about the limiting behaviour of

$$\sup_{\theta\in\Theta} \ln\left(\frac{1}{n} \sum \ln(f_\theta(X_i)\right).$$

This problem would go away if one could establish that the convergence in (9) is *uniform* in θ. This strategy was first realized in a path breaking paper by A. Wald (Wald 1949), where he first established the consistency of the maximum likelihood estimator. Today the techniques are a little more sophisticated. Nevertheless,

consistency proofs for M-estimators still rely to good extend on Wald's idea.

Another application of uniform LLN is the consistency of 'plug-in' estimators. In many cases, the asymptotic variance of certain estimators can be expressed as a function of the expectations of certain random functions, possibly depending on the parameter to be estimated (for example, the well-known 'sandwich formula' derived by H. White; see for example Hayashi 2000). A standard strategy is to estimate the parameter, then replace the expectation by an average (and hope that – due to the LLN – average and expectation are close together) and use the estimated parameter as an argument. One can easily see that only a uniform law of large numbers can justify procedures of this type.

Fortunately, there exist a lot of criteria to establish uniform laws of large numbers. For most cases of interest to econometricians, the papers by Andrews (1992) and Pötscher and Prucha (1989) will be sufficient.

A more general and abstract theory can be found in van der Vaart and Wellner (1996). These theories allow us also to estimate the cumulative distribution function of random variables directly. Suppose we have given random variables X_i,... X_n. Then the empirical distribution function F_n is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

(that is F_n jumps $1/n$ in X_i and is constant in between the jumps). Then the theorem of Glivenko-Cantelli (see van der Vaart and Wellner 1996) states that if the X_i are i.i.d. with cumulative distribution function F, then

$$\sup|F_n(x) - F(x)| \to 0.$$

It should be noted that there are generalizations to multivariate or even more general X_i. In these cases, however, one has to use slightly more sophisticated techniques. Instead of the 'empirical distribution function', one has to use the 'empirical measure' (a random measure, which puts mass $1/n$ in the points X_i, and instead of the maximum difference of the distribution functions one has to consider the maximal difference of the measures over certain classes ('VC-classes').

Bibliography

Andrews, D.W.K. 1992. Generic uniform convergence. *Econometric Theory* 8: 241–257.

Billingsley, P. 1995. *Probability and measure*. 3rd ed. New York: Wiley.

Gray, R.M. 2007. *Probability, random processes, and ergodic properties*. Online. Available at http://ee.stanford.edu/Bgray/arp.html. Accessed 29 Apr 2007.

Hall, P., and C.C. Heyde. 1980. *Martingale limit theory and its application*. San Diego: Academic Press.

Hayashi, F. 2000. *Econometrics*. Princeton: Princeton University Press.

Pötscher, B.M., and I.R. Prucha. 1989. A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica* 57: 675–683.

van der Vaart, A.W., and J.A. Wellner. 1996. *Weak convergence and empirical processes*. New York: Springer.

Wald, A. 1949. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* 20: 595–601.

Law, Economic Analysis of

A. Mitchell Polinsky and Steven Shavell

Abstract

This article surveys the economic analysis of five primary fields of law: property law; liability for accidents; contract law; litigation; and public enforcement and criminal law. It also briefly considers some criticisms of the economic analysis of law.

Keywords

Adverse possession; Asymmetric information; Becker, G.; Bentham, J.; Bona fide purchase rule; Coase, R.; Collective action; Compensated takings; Contract formation; Contractual interpretation; Copyright; Corrective tax; Criminal law; Damage measures; Demsetz,

H.; Deterrence; Direct regulation; Disclosure of information; Due care; Eminent domain; Expectation measure of damages; Fairness; Fault-based liability; Finders-keepers rule; Fines; First-party insurance; Good title; Hold-out problem; Imprisonment; Incapacitation; Incomplete contracts; Injunction; Justice; Labels; Land registries; Law and economics; Liability; Liability insurance; Litigation; Moral hazard; Negligence rule; Original ownership rule; Patents; Posner, R.; Possessory rights; Product liability; Property rights; Public enforcement of law; Public property; Punishment; Risk; Risk aversion; Risk-bearing; Settlement vs trial; Social norms; Social welfare; Specific performance; Strict liability; Suit; Takings; Tort law; Trade secret law; Trademarks; Unilateral accident model; Utilitarianism

JEL Classifications
D02; D23; D63; H23; J28; K11; K12; K13; K14; K32; K41; K42; P14; P48

Economic analysis of law seeks to identify the effects of legal rules on the behaviour of relevant actors and to determine whether these effects are socially desirable. The approach employed is that of economic analysis generally: the behaviour of individuals and firms is described on the assumption that they are forward looking and rational, and the framework of welfare economics is adopted to assess the social desirability of outcomes. The field may be said to have begun with Bentham (1789), who systematically examined how actors would behave in the face of legal incentives (especially criminal sanctions) and who evaluated outcomes with respect to a clearly stated measure of social welfare (utilitarianism). His work was left essentially undeveloped until four important contributions were made: Coase (1960) on externalities and liability, Becker (1968) on crime and law enforcement, Calabresi (1970) on accident law, and Posner (1972) on economic analysis of law in general. (Calabresi's book was the culmination of a series of articles, the first of which was published in 1961; see Calabresi 1961.)

Our focus here is on the analytical foundations of five basic legal subjects: property, torts, contracts, civil litigation, and crime and law enforcement (on these, see generally Cooter and Ulen 2003; Posner 2003; Miceli 1997; and Shavell 2004). We do not treat more particular areas of law, such as antitrust, corporate and tax law, nor do we cite empirical work; for surveys of these and other areas of law and economics, including empirical studies, see Polinsky and Shavell (2007).

Property Law

Justification and Emergence of Property Rights

A beginning question is why there should be property rights in things. A number of arguments have been stressed, especially by early writers, including that property rights furnish incentives to work and to maintain durable things; that the rights make trade possible; and that, if such rights were absent, individuals would spend effort trying to take things from each other and protecting their things.

Property rights would be expected to emerge when their advantages become sufficiently great. For example, Demsetz (1967) explains the development of property rights in land among Indians as a way of preventing overly intensive hunting of valuable animals. Umbeck (1981) shows that when gold was discovered in California in 1848 property rights in gold-bearing land and river beds developed, as this encouraged individuals to pan for gold and to build sluices; it also curbed wasteful efforts to grab land from others. For a survey, see Libecap (1986).

Division of Property Rights

Property rights can be viewed as composed of *possessory* rights – rights of use – and rights to *transfer* possessory rights. Thus, what we commonly conceive of as ownership (say, of land) entails both a large swath of possessory rights

(rights to build on land, plant on it, under most contingencies, and into the infinite future) and associated rights to transfer them. Property rights in things are generally held in substantially agglomerated bundles, but there is also significant partitioning of rights contemporaneously, over time and contingencies, and according to whether the rights are possessory or are for transfer. For example, an owner of land may not hold complete possessory rights, in that others may possess an easement giving them the right of passage upon his land, or the right to take timber, or the right to extract oil if found (thus a contingent right). A rental agreement constitutes a division of property rights over time. Trust arrangements, such as those under which an adult manages property for a child, divide possessory rights and rights to transfer.

The division of property rights may be valuable when different parties derive different benefits from them, because gains can then be achieved if rights are allocated to those who obtain the most from them. There may, however, be disadvantages to the division of rights, including that externalities may arise (a person with a right of passage might trample crops).

Public Property and Its Acquisition; Takings and Compensation

An important class of property is that owned by the public. As is well known, the main justification for public property concerns the difficulty that private providers would experience in charging for certain goods and services.

When it is desirable for the state to acquire property for public use, the state can either purchase it or take it through the exercise of the power of *eminent domain.* In the latter case, the law typically provides that the state must compensate property owners for the value of what has been taken from them.

A difference between purchase and compensated takings is that the amounts owners receive are determined by negotiation in the former case but unilaterally by the state in the latter. Because of errors in state determination of value, as well as concern about the behaviour of government officials, purchase would ordinarily be superior to compensated takings. When, however, the state needs to assemble many contiguous parcels, such as for a road, acquisition by purchase might be stymied by hold-out problems, making the power to take socially advantageous.

On the assumption that there is a reason for the state to take property, a requirement to pay compensation may curb problems of overzealousness or abuse of authority by public officials, yet it may also exacerbate potential problems of insufficient public activity, because public authorities do not directly receive the benefits of takings (Kaplow 1986). Payment of compensation also may lead property owners to invest excessively in property (see Blume et al. 1984).

Acquisition of Property in Unowned Things

The law must determine the conditions under which a person will become a legal owner of previously unowned things, such as wild animals, fish, and mineral and oil deposits. Under the *finders-keepers rule*, incentives to invest in capture (such as to hunt for animals or explore for oil) are optimal if only one person is making the effort. However, if many individuals seek unowned things, they will invest a socially excessive amount of resources in search: one person's investment usually will come, at least partly, at the expense of other person's likelihood of finding unowned things. Various aspects of the law ameliorate this problem of excessive search effort. For example, regulations may limit the quantities of fish and wild animals that can be taken; the right to search for minerals on the ocean floor may be auctioned; and oil extraction rights may be assigned to a single party.

Acquisition of Good Title When Property is Sold

A basic difficulty associated with sale of property that a legal system must solve is establishing validity of ownership or *title*. Good title is important for trade, since buyers want to be assured that they have property rights in what they purchase. But, if any sale gives a buyer good title, theft is encouraged, since thieves could then easily sell stolen goods. Under a *registration system*, good title means that one's name is listed in the registry

as the owner, and title passes at the time of sale by an authorized change in the registry. Hence, buyers can clearly determine whether they are obtaining good title by checking the registry, and a thief could not easily sell stolen property by claiming that he has good title. Registries, however, are expensive to establish and maintain.

In the absence of registries, the law may employ the *original ownership rule,* under which the buyer does not obtain good title if the seller did not have good title. Alternatively, under the *bona fide purchase rule,* a buyer acquires good title as long as he had reason to think that the sale was legitimate, even if the item sold was in fact wrongfully obtained. This rule makes theft more attractive because thieves will often be able to sell their property to buyers who will be motivated to 'believe' that sales are bona fide.

Adverse Possession

The legal doctrine of adverse possession allows involuntary transfer of land: a person is deemed to become the legal owner of land if he takes possession of it and uses it openly for at least a prescribed period, such as ten years. It may appear that this rule could be desirable because it encourages productive use of idle land. But this overlooks the possibility that a prospective adverse possessor could always bargain with the owner to rent or buy the land, and that there may be good reasons for allowing the land to remain idle. Additionally, the rule induces owners to expend resources policing incursions, and potential adverse possessors to attempt possession. A historical justification for the rule is that, before reliable land registries existed, it allowed a seller of land to establish good title to a buyer relatively easily: the seller need only show that he was on the land for the prescribed period.

Constraints on Sale of Property

Legal restrictions are often imposed on the sale of goods and services. One standard justification is externalities. For example, the sale of fireworks might be banned because of the externality that their ownership creates, namely, putting others at risk of injury. The other standard justification for legal restrictions on sale is lack of consumer information. For instance, a drug may not be sold without a prescription because of fear that otherwise buyers would not use it properly. Rather than restrict sales, however, the government could supply relevant information to consumers, such as by indicating that the drug has dangerous side effects, or that it should be taken only on the advice of a medical expert.

Externalities

When individuals use property, they may cause externalities, namely, harm or benefit to others. Generally, it is socially desirable for individuals to do more than is in their self-interest to reduce detrimental externalities and to act so as to increase beneficial externalities. The socially optimal resolution of harmful externalities often involves the behaviour of victims as well as that of injurers. If victims can do things to reduce the amount of harm more cheaply than injurers (say, install air filters to avoid pollution), it is optimal for victims to do so. Moreover, victims can sometimes alter their locations to reduce their exposure to harm.

Legal intervention can ameliorate problems of externalities. A major form of intervention that has been studied is *direct regulation*, under which the state restricts permissible behaviour, such as requiring factories to use smoke arrestors. Closely related is the *injunction*, whereby a potential victim can enlist the power of the state to force a potential injurer to take steps to prevent harm or to cease his activity. Society can also make use of financial incentives to induce injurers to reduce harmful externalities. Under the *corrective tax,* a party pays the state an amount equal to the expected harm he causes – for example, the expected harm due to a discharge of a pollutant into a lake. There is also *liability,* a privately initiated means of providing financial incentives, under which injurers pay for harm done if sued by victims. These methods differ in the information that the state needs to apply them, in whether they require or harness information that victims have about harm, and in other respects, such that each may be superior to the other in different circumstances (Shavell 1993).

Parties affected by externalities will sometimes have the opportunity to make mutually beneficial

agreements with those who generate the externalities, as Coase (1960) stressed. But bargaining may not occur, for many reasons: cost; collective action problems (such as when many victims each face small harms); and lack of knowledge of harm (such as from an invisible carcinogen). If bargaining does occur, it may not be successful, owing to asymmetric information. These difficulties often make bargaining a problematic solution to externality problems and imply that liability rules are needed, as discussed by Calabresi and Melamed (1972).

Property Rights in Information

The granting of property rights in information, notably the award of *patents* for inventions and *copyrights* for written works and certain other compositions, involves a major social benefit – the provision of incentives to create intellectual works – but also a social disadvantage – the creation of power to price above marginal cost. Patent and copyright law have been examined to ascertain how they reflect the tradeoff between this benefit and disadvantage. A distinct form of legal protection is *trade secret law,* comprising various doctrines of contract and tort law that serve to protect a range of commercially valuable information that is not (or cannot be) protected by patent or copyright, such as customer lists. On property rights in information, see generally Landes and Posner (2003).

An alternative to property rights in information is for the state to offer *rewards* to creators of information, and for information that is developed to be made available to all who want it. Thus, an author of a book would receive a reward from the state for writing the book, possibly based on sales, but anyone who wanted to print it and sell it could do so. This system would create incentives for the creation of information without distorting prices, but requires the state to choose the magnitude of rewards.

Property Rights in Labels

Many goods and services are identified by labels, which have substantial social value because the quality of goods and services may be hard for consumers to determine directly. Labels enable consumers to purchase goods and services on the basis of product quality without requiring consumers to independently determine quality; a person who wants to stay at a high-quality hotel in another city can choose such a hotel merely by its label, such as 'Ritz Hotel'. In addition, sellers who label their output will have an incentive to produce goods and services of quality because consumers will recognize quality through sellers' labels. This basic reasoning is used to justify property rights in *trademarks,* as discussed by Landes and Posner (1987b).

Liability for Accidents

Legal liability for accidents, which is governed by tort law, is a means by which society can reduce the risk of harm by threatening potential injurers with having to pay for the harms they cause. Liability is also frequently viewed as a device for compensating victims of harm, though we emphasize that insurance can provide compensation more cheaply than the liability system. There are two basic rules of liability. Under *strict liability,* an injurer must always pay for harm due to an accident that he causes. Under the *negligence rule,* an injurer must pay for harm caused only when he is found negligent, that is, only when his level of care was less than a standard of care chosen by the courts, often referred to as *due care.* (There are various versions of these rules that depend on whether victims' care was insufficient.) In practice, the negligence rule is the dominant form of liability; strict liability is reserved mainly for certain especially dangerous activities. On economic analysis of liability for accidents, see generally Calabresi (1970), Landes and Posner (1987a), and Shavell (1987a).

Incentives to Take Care

In order to focus on how liability affects the incentive to prevent harm, assume first that parties are risk neutral and that accidents are *unilateral* – only injurers (not victims) influence risk by their choice of *care x.* Let $p(x)$ be the probability of an accident that causes harm h, where p is declining in x. Assume that the social objective is to

minimize total expected costs, $x + p(x)h$, and let x^* denote the optimal x.

Under strict liability, injurers pay damages equal to h whenever an accident occurs, and they naturally bear the cost of care x. Thus, they minimize $x + p(x)h$; accordingly, they choose x^*.

Under the negligence rule, suppose that the due care level is set equal to x^*, meaning that an injurer who causes harm will have to pay h if $x < x^*$, but will not have to pay anything if $x \geq x^*$. Then it can be shown that the injurer will choose x^*: clearly, the injurer will not choose x greater than x^*; and he will not choose $x < x^*$, for then he will be liable (in which case the analysis of strict liability shows that he would not choose $x < x^*$). Thus, under both forms of liability, injurers are led to take optimal care. Note that to apply the negligence rule courts need sufficient information to calculate x^* and to observe x, whereas under strict liability they only have to observe x.

The analysis of incentives and liability has been undertaken as well for *bilateral* accidents, in which victims also take care, and when there is uncertainty in the determination of negligence (such as due to imperfect observation of x). On incentives and liability for unilateral and bilateral accidents, see originally Brown (1973) and also Diamond (1974).

Level of Activity

An important extension allows for injurers to choose their *level of activity* z, which is interpreted as the (continuously variable) number of times they engage in their activity (or, if injurers are firms, their output). Let $b(z)$ be the benefit (or profit) from the activity, and assume the social objective is to maximize $b(z) - z(x + p(x)h)$; here $x + p(x)h$ is assumed to be the cost of care and expected harm each time an injurer engages in his activity. Let x^* and z^* be optimal values. Note that x^* minimizes $x + p(x)h$, so x^* is as described above, and that z^* is determined by $b'(z) = x^* + p(x^*)h$, which is to say, the marginal benefit from the activity equals the marginal social cost.

Under strict liability, an injurer will choose both the level of care and the level of activity optimally, as his objective will be the same as the social objective, to maximize $b(z) - z(x + p(x)h)$, because damage payments equal h whenever harm occurs. Under the negligence rule, an injurer will choose optimal care x^* as before, but his level of activity z will be socially excessive. In particular, because an injurer will escape liability by taking care x^*, he will choose z to maximize $b(z) - zx^*$, so that z will satisfy $b'(z) = x^*$. The injurer's cost of raising his level of activity is only his cost of care x^*, which is less than the social cost, which also includes $p(x^*)h$. On liability and the level of activity, see Shavell (1980b).

The failure of the negligence rule to control the level of activity arises because negligence is defined here (and also generally in practice) in terms of care alone. A justification for this restriction is the difficulty courts would face in determining the optimal activity level z^* and the actual z. The failure of the negligence rule to control the injurer's level of activity is applicable to any aspect of injurer behaviour that would be difficult to regulate directly (including, for example, research and development activity). If, however, courts were able to incorporate all aspects of injurer behaviour into the definition of due care, the negligence rule would result in optimal behaviour in all respects. (Note that the variable x in the original problem could be interpreted as a vector, with each element corresponding to a dimension of behaviour.)

Product Liability

Another extension of the model of liability and incentives concerns product liability, the liability of firms for harms suffered by their customers. Here the degree to which liability creates incentives to reduce risk depends on customer knowledge of risk. If their knowledge is perfect, liability does not affect incentives since customers will recognize risky products and pay appropriately less for them. If their knowledge is imperfect, there is a role for liability, in many respects similar to what has been discussed above.

Risk-Bearing and Insurance

In addition to affecting incentives to reduce harm, the socially optimal resolution of the accident problem involves the spreading of risk to lessen

risk-bearing by risk-averse parties. Risk-bearing is relevant not only because potential victims may face the risk of accident losses, but also because potential injurers may face the risk of liability. The former risk can be mitigated through so-called first-party insurance that covers losses suffered in accidents, and the latter through liability insurance.

Because risk-averse individuals tend to purchase insurance, the incentives associated with liability do not function in the direct way discussed above, but instead are mediated by the terms of insurance policies. To illustrate, consider strict liability in the unilateral accident model with care alone allowed to vary, and assume that insurance is sold at actuarially fair rates. If injurers are risk averse and liability insurers can observe their levels of care, injurers will purchase full liability insurance coverage and their premiums will depend on their level of care; their premiums will equal $p(x)h$. Thus, injurers will want to minimize their costs of care plus premiums, or $x + p(x)h$, so they will choose the optimal level of care x^*. In this instance, liability insurance eliminates risk for injurers, and the situation reduces to the previously analysed risk-neutral case. (Victims do not bear risk either because, in the present case, they are fully compensated for their losses.)

If, however, liability insurers cannot observe levels of care, insurance policies with full coverage could create severe moral hazard, and so might not be purchased. Instead, as we know from the theory of insurance, the typical amount of coverage purchased will be partial, for that leaves injurers with an incentive to reduce risk. In this case, therefore, the liability rule results in some direct incentive to take care because injurers are left bearing some risk after their purchase of liability insurance. But levels of care will still tend to be less than first-best.

This last observation raises the question of whether the sale of liability insurance is socially desirable. (We note that because of concern about diluted incentives, liability insurance was delayed for decades in many countries and is sometimes forbidden today, such as for punitive damages.) Notwithstanding the moral hazard problem, the sale of liability insurance is socially desirable, at least in basic models of accidents and some variations of them. This is because, if the liability insurer and the injurer together have to pay for the harm caused, the insurance policy will appropriately balance the social desire to reduce harm and the social desire to reduce risk-bearing.

Parallel observations apply under the negligence rule, where the focus of concern is on the bearing of risk by victims since injurers generally will take due care and not be liable. Risk-averse potential victims will tend to purchase first-party accident insurance.

The presence of insurance implies that the liability system cannot be justified primarily as a means of compensating risk-averse victims against loss. Rather, the justification for the liability system must lie in significant part in the incentives that it creates to reduce risk. To amplify, although both strict liability and the insurance system can compensate victims, the liability system is much more expensive than the insurance system (see below). Accordingly, if there were not a social need to create incentives to reduce risk, it would be best to dispense with the liability system and to rely on insurance to accomplish compensation. On liability and insurance, see Shavell (1982a).

Administrative Costs

The administrative costs of the liability system – the legal costs and effort of litigants involved in suit, settlement and trial – are substantial, generally exceeding the amounts received by victims. Consideration of administrative costs affects the comparison of liability rules, but it is not clear which rule involves greater expense: more cases are brought under strict liability than under the negligence rule (victims will not sue under the negligence rule if they believe the injurer was not negligent), but the cost of resolving a case should be greater under the negligence rule (because due care and the injurer's care level need to be ascertained). The presence of administrative costs raises the questions of whether the incentive benefits of the liability system justify incurring these costs, and whether the private incentive to sue is socially optimal. These questions are discussed in Sect. "Litigation".

Contracts

A contract is a specification of the actions that named parties are supposed to take at various times, as a function of the conditions that then obtain. A contract is said to be completely detailed, or simply *complete,* if the contract provides *explicitly* for all possible conditions. An incomplete contract may well cover all conditions by implication. A contract stating merely that a specified price will be paid for a bushel of wheat is incomplete because it does not mention many contingencies that might affect the parties. Note that such an incomplete contract has no *gaps,* as it stipulates what the parties are to do in all circumstances. Typically, incomplete contracts do not include conditions which, were they easy to include, would allow both parties to be made better off in an expected sense.

Contracts are here assumed to be enforced by a tribunal, which will usually be interpreted to be a state-authorized court, but it could also be another entity, such as an arbitrator or the decision-making body of a trade association or a religious group. (Reputation and other non-legal factors may also serve to enforce contracts, but we do not discuss these.) Enforcement refers to actions taken by the tribunal when one or more of the parties to the contract decide to come before it.

General Reasons for Contracts

Broadly speaking, parties make contracts when they have a need to make plans.

They also want contracts enforced to prevent opportunistic behaviour that otherwise might occur during the course of the contractual relationship and stymie fulfilment of their plans.

There are two basic contexts in which parties make enforceable contracts. The first concerns virtually any kind of financial arrangement. The necessity of contract enforcement here is transparent. In financial arrangements, there is often a party who extends credit to another for some time period, and contract enforcement prevents his credit from being appropriated, which otherwise would render the arrangements impossible. For example, if borrowers were not forced to repay loans, loans would be unworkable. In addition, financial contracts that allocate risk would generally be useless without enforcement because, once the risky outcome became known, one of the parties would not wish to honour the contract.

The second context in which parties make enforceable contracts involves the supply of customized or specialized goods and services which cannot be purchased on a spot market with a simultaneous exchange for money. The need for enforcement of agreements for supply of customized goods and services inheres in several advantages: averting problems of hold-up, which might distort incentives to invest in the contractual enterprise; allocation of risk; and prevention of inappropriate breach or performance, which can result from imperfect bargaining due to sheer cost or asymmetric information.

Contract Formation

The formation of contracts is of interest, in several respects. One issue concerns search effort (Diamond and Maskin 1979). Parties expend effort in finding contractual partners, and it is apparent that their search effort will not generally be socially optimal. On the one hand, they might not search enough: because the joint gain from contracting will generally be divided between the parties through the bargaining process, the private return to search may be less than the social return. On the other hand, parties might search more than is socially desirable because of a negative externality associated with discovery of a contract partner: when one party finds and contracts with a second, other parties are thereby prevented from contracting with that party.

A basic question that a tribunal must answer is: at what stage of interactions between parties does a contract become legally recognized? The general legal rule is that contracts are recognized if and only if both parties give a clear indication of assent, such as signing their names on a document. This rule allows parties to make enforceable contracts when they so desire, and it also protects parties from becoming legally obliged against their wishes, such as from one party's reliance on the other's statements (Bebchuk and Ben-Shahar 2001; Wils 1993). Mutual assent

sometimes is not simultaneous; one party will make an offer and time will pass before the other agrees. An issue that this raises is how long, and under what circumstances, the offeror will want to be held to his offer, and whether he should be held to it. If an offeror is held to his terms, offerees will often be led to invest effort in investigating contractual opportunities. Otherwise, offerees might be taken advantage of by offerors if the offerees expressed serious interest after costly investigation (the offeror could change to less favourable terms). The anticipation of such offeror advantage-taking would reduce offerees' incentive to engage in investigation and thus diminish mutually beneficial contract formation (see, for example, Craswell 1996; Katz 1990, 1996).

Another issue of note is disclosure of information at the time of contract formation. Disclosure may be socially beneficial because the disclosed information may be desirably employed by one of the parties; for example, a buyer of a house may learn from the seller that the basement leaks and thus decide not to store valuables there. However, a disclosure obligation discourages parties from investing in acquisition of information (Kronman 1978). For instance, an oil company contemplating buying land might decide against conducting a geological analysis of it to determine its oil-bearing potential if the company would be required to disclose its findings to the seller of the land, as the seller would then demand a price reflecting the value of the land. The social welfare consequences of the effect of a disclosure obligation on the motive to acquire information depend on whether the information is socially valuable or mere foreknowledge, on whether the party acquiring information is the buyer or the seller, and on inferences that would be drawn from silence (Shavell 1994).

Even if both parties have given their assent, a contract will not be recognized if it was made when one of the parties was put under undue pressure – for example, if a party was physically or otherwise threatened by another. This legal rule has virtues similar to those of laws against theft; it reduces individuals' incentives to expend effort making threats and defending themselves against threats.

In addition, contracts may not be legally recognized if they are made in emergency situations, such as when the owner of a ship in distress promises to pay an exorbitant amount for rescue. Non-enforcement in such situations beneficially provides potential victims with implicit insurance against having to pay high prices, but it also reduces incentives for rescue.

Incomplete Nature of Contracts and Their Less-Than-Rigorous Enforcement

Contracts are commonly observed to be significantly incomplete, leaving out all manner of variables and contingencies that are of potential relevance to contracting parties. Moreover, contracts are not enforced with high sanctions, and breach is not an uncommon event.

There are three reasons for the incompleteness of contracts. The first is the cost of writing more complete contracts. The second is that some variables (effort levels, technical production difficulties) cannot be verified by tribunals. The third is that the expected consequences of incompleteness may not be very harmful to contracting parties. Incompleteness may not be harmful because a tribunal might interpret an imperfect contract in a desirable manner. Also, as will be seen, the prospect of having to pay damages for breach of contract may serve as an implicit substitute for more detailed terms. Furthermore, the opportunity to renegotiate a contract often furnishes a way for parties to alter terms in the light of circumstances for which contractual provisions had not been made.

Interpretation of Contracts

Contractual interpretation, which includes a tribunal's filling gaps, resolving ambiguities, and overriding literal language, can benefit parties by easing their drafting burdens or reducing their need to understand contractual detail. For example, if it is efficient to excuse a seller from having to perform if his factory burns down, the parties need not incur the cost of specifying this exception in their contract if they can trust the tribunal to interpret their contract as if the exception were specified. A method of interpretation can be viewed formally as a function that transforms the

contract individuals write into the effective contract that the tribunal will enforce. Given a method of interpretation, parties will choose contracts in a constrained-efficient way. Notably, if the parties are concerned that an aspect of their contract would not be interpreted as they want, they could either bear the cost of writing a more explicit term that would be respected by the tribunal, or they could simply accept the expected loss from having a less-than-efficient term. The socially optimal method of interpretation will take this reaction of contracting parties into account and can be regarded as minimizing the sum of the costs the parties bear in writing contracts and the losses resulting from inefficient enforcement. (See Ayres and Gertner 1989; Hadfield 1994; Schwartz 1992; Shavell 2006.)

Damage Measures for Breach of Contract

When parties breach a contract, they often have to pay damages in consequence. The damage measure, the formula governing what they should pay, can be determined by the tribunal or it can be stipulated in advance by the parties to the contract. One would expect parties to specify their own damage measure when it would better serve their purposes than the measure the tribunal would employ, and otherwise to allow the tribunal to select the damage measure. In either case, we now examine the utility of different damage measures to contracting parties, assuming initially that there is no renegotiation of contracts.

Clearly, the prospect of having to pay damages provides an incentive to perform contractual obligations, and thus generally promotes enforcement of contracts and the goals of the parties. Under the commonly employed *expectation measure,* damages equal the amount that compensates the victim of breach for his losses. Under this measure, a seller contemplating breach will be induced to perform if the cost of performance to the seller is less than the value of performance to the buyer, and to breach otherwise. Because the expectation measure leads to maximization of joint value, it would be chosen by the parties (ignoring consideration of investment incentives and risk bearing), as emphasized by Shavell (1980a). Another commonly employed measure of damages is the *reliance measure:* damages equal to the amount spent by the victim relying on contract performance, such as expenditures on advertising an entertainer who has contracted to appear at one's nightclub.

The point that the expectation measure of damages induces efficient performance of parties sheds light on the view of many legal commentators that breach is immoral. This view fails to account for the fact that contracts that are breached are generally incomplete, and that breach constitutes behaviour that the parties truly want and would have provided for in a complete contract.

Damage measures not only affect performance, they also influence the *ex ante* motive to make investments in reliance on contract performance. Under the expectation measure, reliance investments tend to exceed efficient levels: the buyer will treat an investment (like advertising an entertainer) as one with a sure payoff, since he will receive either performance or expectation damages, whereas the actual return to the investment is uncertain, due to the possibility of breach (advertising will be a waste if the entertainer does not appear); see Shavell (1980a). This tendency toward over-reliance stands in contrast to the problem of inadequate reliance investment associated with lack of contract enforcement.

Damage measures affect risk-bearing as well as incentives. Notably, because the expectation measure compensates the victim of a breach, the measure might be mutually desirable as a form of insurance if the victim is risk averse (Polinsky 1983). However, the prospect of having to pay damages also constitutes a risk for a party who might commit breach (such as a seller whose costs suddenly rise), and he might be risk averse as well. The latter consideration may lead parties to want to lower damages or to employ damages less frequently by writing more detailed contracts (for instance, the parties could go to the expense of specifying in the contract that a seller can be excused from performance if his costs are unusually high).

Specific Performance as a Remedy for Breach

An alternative to use of a damage measure for breach of contract is specific performance:

requiring a party to satisfy his contractual obligation. Specific performance can be accomplished with a sufficiently high threat or by exercise of the state's police powers, such as by a sheriff removing a person from the land that he promised to convey. (Note that, if a monetary penalty can be employed to induce performance, then specific performance is equivalent to a damage measure with a high level of damages.)

It is apparent from what has been said about incomplete contracts and damage measures that parties should not want specific performance of many contracts that they write, for they do not wish their incomplete contracts always to be performed. It is therefore not surprising that, in fact, specific performance is not used as the remedy for breach for most contracts for production of goods or for provision of services. Additionally, specific performance might be peculiarly difficult to enforce in these contexts because of problems in monitoring and controlling parties' effort levels and the quality of production.

However, specific performance does have advantages for parties in certain contexts, such as in contracts for the transfer of things that already exist, like land, and specific performance is the usual legal remedy for sellers' breaches of contracts for the sale of land.

Renegotiation of Contracts

Parties often have the opportunity to renegotiate their contracts when problems arise. Indeed, the assumption that they will do this has appeal because, having made an initial contract, the parties know of each other's existence and of many particulars of the contractual situation. For this reason, much of the economics literature (as opposed to law and economics literature) on contracts assumes that renegotiation always occurs and that, due to symmetric information between the parties, it always results in efficient performance. Hence, damage measures for breach of contract, or more generally, the mechanisms that the parties stipulate in their contracts, establish the threat points for renegotiation. If properly designed, the mechanisms can foster beneficial incentives to invest *ex ante* for both parties. On this extensive literature, see, for example, Rogerson (1984), Hart (1987), Hart and Moore (1988), and Bolton and Dewatripont (2005).

Legal Overriding of Contracts

A basic rationale for legislative or judicial overriding of contracts is the presence of externalities. Contracts that are likely to harm third parties are often not enforced, including, for example, agreements to commit crimes, price-fixing compacts, liability insurance policies against fines, and certain sales contracts (such as for machine guns).

Another general rationale for non-enforcement of contracts is to prevent a loss in welfare to one or both of the parties to a contract. This concern may justify nonenforcement when a party is incompetent, lacks relevant information, or is in an emergency situation. The rationale also applies in the context of contract interpretation by tribunals. As noted, contract interpretation may amount to the overriding of a written contractual term, and this practice may promote the welfare of contracting parties by allowing them to save writing costs, given that courts will step in and correct inefficient terms.

Additionally, contracts sometimes are not enforced because they involve the sale of things said to be inalienable, such as human organs, babies, and voting rights. In many of these cases, the inalienability justification for lack of enforcement can be recognized as involving externalities or the welfare of the contracting parties.

Litigation

We here consider the bringing and adjudication of lawsuits: the decision of a party who has suffered a loss whether to sue; the choice of the litigants whether to settle with each other or instead go to trial; and the choice of litigants, before or during trial, of how much to spend on litigation.

Suit

As a general rule, a party who has suffered loss, the plaintiff, will sue when the cost of suit c_P is less than the expected benefits from suit. The expected benefits from suit incorporate potential settlements or trial outcomes, but assume for

simplicity that, if suit is brought, the plaintiff obtains for sure a judgment equal to harm suffered, h. Thus the plaintiff will sue when his litigation cost, c_P is less than h. (Obviously, if there is only a probability p of winning this amount, a risk-neutral plaintiff would sue when $c_P < ph$; and a risk-averse plaintiff would be less likely to sue.)

The private incentive to sue is fundamentally misaligned with the socially optimal incentive to sue, as emphasized by Shavell (1982b, 1997). The deviation could be in either direction. On the one hand, there is a divergence between private and social costs that can lead to socially excessive suit: when a plaintiff contemplates bringing suit, he bears only his own costs; he does not take into account the defendant's costs or the state's costs that his suit will engender. On the other hand, there is a difference between the private and social benefits of suit that can either lead to a socially inadequate level of suit or reinforce the cost-related tendency towards excessive suit. Specifically, the plaintiff considers his private benefit from suit (the gain he would obtain from prevailing) but not the social benefit (the deterrent effect on the behaviour of injurers generally). The private gain could be larger or smaller than the social benefit.

To illustrate, suppose that liability is strict. As stated, victims will sue if and only if $c_P < h$. Let x be the precaution expenditures that injurers will be induced to make if there is suit, q the probability of harm if suit is not brought, and q' the probability of harm if suit is brought. (Thus, q' will be less than q if x is spent on precautions.) Suit will be socially worthwhile if and only if $q'(c_P + c_D + c_S) < (q - q')\,h - x$, where c_D is the defendant's litigation cost and c_S is the state's cost. In other words, suit is socially worthwhile if the expected litigation costs are less than the deterrence benefits of suit net of the cost of precautions. The condition for victims to sue and the condition for suit to be socially optimal are very different. Whether victims will sue does not depend on the costs c_D and c_S. Moreover, the private benefit of suit is what the victim will receive as a damages award, h; in contrast, the social benefit is the harm weighted by the reduction in the accident probability, $q - q'$, net of the cost of precautions, x. It is evident, therefore, that victims might sue when suit is not socially desirable, or that victims might not sue even when suit would be socially beneficial.

The main implication of the private-social divergence is that state intervention may be desirable, either to correct a problem of excessive suit (notably by taxing suit or barring it in some domain) or a problem of inadequate suit (by subsidizing suit in some way). For the state to determine optimal policy, however, requires it to estimate the effects of suit on injurer behaviour and weigh them against the social costs of suit.

The importance of the private-social divergence in incentives to sue may be substantial. This is suggested by the high costs of using the legal system; indeed, legal costs may on average actually equal the amounts received by those who sue. Hence, the incentives created by the legal system must be significant to justify its use. Regardless of whether the legal system creates valuable incentives, however, the private motive to bring suit may be great, giving rise to a reason for social intervention. Conversely, in some domains the incentive to sue may be low (say, damages per plaintiff are not great) even though the value of deterrence is significant. This might justify the state's encouraging litigation.

Settlement Versus Trial

Assuming that a suit has been brought, we now consider whether parties will reach a settlement or go to trial. A settlement is a legally enforceable contract, usually involving a payment from the defendant to the plaintiff, in return for which the plaintiff agrees not to pursue his claim further. If the parties do not reach a settlement, we assume that they go to trial, that is, that some tribunal determines the outcome of their case. In fact, the vast majority of cases settle.

One model of the settlement-versus-trial decision presumes that the parties have somehow each come to a belief about the probability of the trial outcome (Posner 2003, ch. 21; Shavell 2004, ch. 17). Let p_P represent the plaintiff's opinion about his probability of prevailing, and let p_D be the defendant's opinion about that same

probability. Let w be the amount that would be won (for simplicity assume that they agree about w). Assume also that the parties are risk neutral. The plaintiff's expected gain from trial, net of his litigation costs, is $p_P w - c_P$. The defendant's expected loss from trial, including his litigation costs, is $p_D w + c_D$. Hence, a settlement is possible if and only if $p_P w - c_P > p_D w + c_D$, in which case the settlement amount will be in the settlement range $[p_P w - c_P, p_D w + c_D]$. Note that, if the parties agree on the plaintiff's probability of prevailing, a settlement is feasible. A settlement range does not exist, and therefore trial will occur, if $p_P w - p_D w > c_P + c_D$. Risk aversion of the parties increases the size of the settlement range and thus, one presumes, makes settlement more likely: if the plaintiff is risk averse, he will be willing to settle for less than $p_P w - c_P$; and if the defendant is risk averse, she will be willing to pay more than $p_D w + c_D$.

The model just discussed does not explain the origin of the parties' beliefs and does not include a description of rational bargaining between them. Subsequently, standard asymmetric information models of settlement versus litigation were examined (Bebchuk 1984; Reinganum and Wilde 1986; Schweizer 1989; Spier 1992; Hay and Spier 1998; Daughety 2000). In a simple model of this type, there is one-sided asymmetry of information and the party without private information makes a take-it-or-leave-it settlement proposal. For example, the plaintiff makes a demand x to the defendant, who has private information about the probability p that he will lose at trial. If $pw + c_D < x$, the defendant will reject the demand and the plaintiff will therefore obtain only $pw - c_P$, but if $pw + c_D > x$, the defendant will accept and pay x. The plaintiff chooses x to maximize his expected payoff from settlement or trial. The higher his demand x, the more he will obtain if it is accepted, but the greater the likelihood of rejection and thus of his bearing trial costs. At the optimal demand for the plaintiff, there will generally be a positive probability of trial and also of settlement.

The virtues of such asymmetric information models are twofold. First, they include an explicit account of bargaining and thus of the probability of settlement and the magnitude of the settlement offer or demand. (The outcomes of these models depend, however, on essentially arbitrary modelling choices, such as whether the informed or the uninformed party makes the settlement proposal.) Second, the models explain differences of opinion that give rise to trial in terms of differences in possession of information. (However, the models do not account for why there should be differences in information, given that the parties have incentives to share information and may be forced to do so through legal discovery.)

The private and social incentives to settle generally diverge for several reasons. First, because the litigants do not bear all of the costs of a trial (such as the salaries of judges and the forgone value of juror time), they save less by settling than society does, which tends to make the private incentive to settle socially inadequate. Second, when there is asymmetric information, parties will fail to settle when the plaintiff's demand turns out to have been too high or the defendant's offer too low. But their desire to obtain from each other a greater share of the benefit from settling does not itself translate into any social benefit. Third, the prospect of settlement may reduce deterrence because defendants gain from settlement.

Litigation Expenditures

A plaintiff will continue spending on litigation as long as this raises his expected return from settlement or trial (net of litigation costs), and a defendant will make such expenditures as long as this lowers his expected total outlays. The effects of each litigant's expenditures will generally depend on what the other does, and the two will often be spending to rebut one another.

There are several reasons why the private and social incentives to spend on litigation diverge. First, to the extent that their expenditures simply offset each other, without altering trial or settlement outcomes, the expenditures constitute a social waste. Second, the litigants' trial expenditures may mislead the tribunal rather than enhance the accuracy of the outcome, which has negative social value. Third, even if trial expenditures do improve the accuracy of outcomes, they may not

be socially optimal in magnitude, for the parties consider only how their expenditures influence the litigation outcome, without regard to their influence (if any) on deterrence.

Because private and social incentives to spend on litigation may diverge, it may be beneficial for expenditures to be either curtailed or encouraged. In practice, courts often restrict the legal effort that parties can undertake, for example by limiting the extent of discovery and the number of testifying experts.

Other Topics

A number of other topics that relate to litigation and the legal process have been studied, including the selection of suits for litigation (Priest and Klein 1984); the accuracy of adjudication (Kaplow 1994; Png 1986); 'discovery', that is, mandated disclosure of information during litigation (Shavell 1989); and the appeals process (Daughety and Reinganum 2000; Shavell 1995; Spitzer and Talley 2000).

Public Law Enforcement and Criminal Law

Law enforcement often is the result of the efforts of public agents, such as inspectors, tax auditors, and police. We here discuss certain characteristics of optimal public law enforcement. As noted, this subject was first analysed by Bentham (1789) and Becker (1968) (for a survey, see Polinsky and Shavell 2000).

Rationale of Public Enforcement

A basic question is why there is a need for public enforcement of law in the light of the availability of private suits brought by victims (Becker and Stigler 1974; Landes and Posner 1975; Polinsky 1980). The answer depends importantly on the locus of information about the identity of injurers. When victims of harm naturally possess knowledge of the identity of injurers, allowing private suits for damages will motivate victims to sue and thus harness the information they have for purposes of law enforcement. This may help to explain why the enforcement of contractual obligations and of accident law is primarily private. When victims do not know who caused harm, however, or when finding injurers is difficult, society tends to rely instead on public investigation and prosecution; this is broadly true of crimes and of many violations of environmental and safety regulations.

Basic Framework for Analysing Public Enforcement

Suppose that, if an individual commits a harmful act, he obtains a gain and also faces the risk of being caught and sanctioned. The sanction could be a fine or a prison term. Fines will be treated as socially costless because they are mere transfers of money, whereas imprisonment is socially costly because of the expense of operating prisons and the disutility suffered by those imprisoned (which is not offset by gains to others). The higher the probability is of detecting and sanctioning violators, the more resources the state must devote to enforcement.

We assume that social welfare equals the sum of individuals' expected utilities. If individuals are risk neutral, social welfare can be expressed as the gains individuals obtain from committing their harmful acts, minus the harms caused and the costs of law enforcement. The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures, or, equivalently, a probability of detection, the form of sanctions, and their level.

Fines

Suppose that the sanction is a fine and that individuals are risk neutral. Then the optimal level of the fine is maximal, f_M, as emphasized in Becker (1968). If the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence. Formally, if $f < f_M$, then raise the fine to f_M and lower the probability from p to $(f/f_M)\,p$; the expected fine is still pf, so that deterrence is maintained, but expenditures on enforcement are reduced, implying that social welfare rises. Moreover, the optimal probability is such that there is some under-deterrence; in other words, at the optimal p the expected fine pf_M is

less than the harm h. The reason for this result is that, if pf_M equals h, behaviour will be ideal, in which case decreasing p must be socially beneficial because the individuals thereby induced to commit the harmful act cause no net social losses (because their gains essentially equal the harm), but reducing p saves enforcement costs.

If individuals are risk averse, the optimal fine may well be below the maximal fine, as stressed in Polinsky and Shavell (1979). This is because the use of a very high fine would impose a substantial risk-bearing cost on individuals who commit harmful acts.

Imprisonment

Now suppose that the sanction is imprisonment and that individuals are risk neutral in imprisonment. Then the optimal imprisonment term is maximal. The reasoning is similar to that employed above with respect to fines: if the imprisonment term were not maximal, it could be raised and the probability of detection lowered so as to keep the expected prison term constant; neither individual behaviour nor the costs of imposing imprisonment are affected (because the expected prison term is the same), but enforcement expenditures fall.

If, instead, individuals are risk averse in imprisonment (the disutility of each additional year of imprisonment grows with the number of years in prison), there is a stronger argument for setting the imprisonment sanction maximally than when individuals are risk neutral. Now, when the imprisonment term is raised, the probability of detection can be lowered even more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but the social costs of imposing imprisonment sanctions decline because the expected prison term falls.

Last, suppose that individuals are risk preferring in imprisonment (the disutility of each additional year of imprisonment declines with the number of years in prison). This possibility seems particularly important: the first years of imprisonment may create unusually high disutility, due to brutalization of the prisoner or due to the stigma of having been imprisoned at all. In addition, individuals generally have positive time discount rates, which are thought to be especially significant for criminals. In the case of risk-preferring individuals, the optimal prison term may well be less than maximal: if the sentence were raised, the probability that maintains deterrence could not be lowered proportionally, implying that the expected prison term would rise. Thus, although there would be enforcement-cost savings, they might not be great enough to offset the increased sanctioning costs.

Fines versus imprisonment

Fines generally are preferable to prison terms as a means of deterrence, since fines are socially cheaper sanctions to impose (Becker 1968). Hence, fines should be employed to the greatest extent possible – until a party's wealth is exhausted – before imprisonment is imposed. Further, imprisonment should be used as a sanction only if the harm prevented by the added deterrence is sufficiently great.

Fault-Based Liability

Our discussion so far has presumed that liability is strict, but liability may also be based on fault, an assessment of whether the act that caused harm was socially undesirable (analogous to the negligence rule and due-care standard discussed above in the accident context). Fault-based liability, like strict liability, can induce individuals to behave properly, but fault-based liability possesses an advantage when individuals are risk averse: if they act responsibly, they will not be found at fault, so will not bear the risk of being sanctioned. Similarly, fault-based liability is advantageous when the form of the sanction is imprisonment, for then, again, individuals may be led to behave optimally without the actual imposition of sanctions, and thus without social costs being incurred (Shavell 1987b). To the extent that mistakes are made in determining fault, however, these two advantages are reduced because risk is imposed and sanctioning costs are incurred. Note, too, that fault-based liability is more difficult to implement, because it requires the state to determine optimal behaviour.

Incapacitation

Society may reduce harm not only through deterrence but also by imposing sanctions that remove parties from positions in which they are able to cause harm, that is, by incapacitating them. Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their drivers' licences, businesses can lose their right to operate in certain domains, and the like.

Suppose that the sole function of imprisonment is to incapacitate. Then it will be desirable to keep someone in jail as long as the reduction in crime from incapacitating him exceeds the costs of imprisonment (Shavell 1987c). Although this condition could hold for a long period, it is unlikely to unless the harm prevented is very high, because the proclivity to commit crimes apparently declines sharply with age.

Note that, as a matter of economic logic, the incapacitation rationale might imply that a person should be imprisoned even if he has not committed a crime – because the danger he poses to society makes incapacitating him worthwhile. In practice, however, the fact that a person has committed a harmful act may be the best basis for predicting his future behaviour, in which case the incapacitation rationale would suggest imprisoning an individual only if he has committed such an act.

Two observations are worth noting about optimal enforcement when incapacitation is the goal as opposed to when deterrence is the goal. First, when enforcement is based on incapacitation, the optimal magnitude of the sanction is independent of the probability of apprehension, which contrasts with the case when enforcement is based on deterrence. Second, when enforcement is deterrence-oriented, the probability and magnitude of sanctions depend on the ability to deter, and, if this ability is limited (as, for instance, with the insane), a low expected sanction may be optimal, whereas a high sanction still might be called for to incapacitate.

Other Issues

A number of other topics have been studied in the economic analysis of public law enforcement, including mistake, marginal deterrence (the effect of sanctions in reducing the severity of harm a party causes), self-reporting of violations (Kaplow and Shavell 1994a; Innes 1999), repeat offences, plea bargaining (Reinganum 1988), general enforcement (when detection resources simultaneously influence the deterrence of a range of harmful acts) (Mookherjee and Png 1992; and Shavell 1991), and corruption of law-enforcement agents (Shleifer and Vishny 1993; Rose-Ackerman 1999; and Polinsky and Shavell 2001).

Criminal Law

The subject of criminal law may be viewed in the light of the theory of public law enforcement (Posner 1985; Shavell 1985). First, the fact that the acts in the core area of crime (robbery, murder, rape, and so forth) are punished by the sanction of imprisonment makes basic sense. Were society to rely on fines alone, deterrence of the acts in question would be grossly inadequate. Notably, the probability of detecting many of these acts is low, making the money sanction necessary for deterrence high, but the assets of individuals who commit these acts often are insubstantial. Hence, the threat of prison is needed for deterrence. Moreover, the incapacitative aspect of imprisonment is valuable because of the difficulty of deterring individuals who are prone to commit criminal acts.

Second, many of the doctrines of criminal law appear to enhance social welfare. This seems true of the basic feature of criminal law that punishment is not imposed on all harmful acts, but instead is usually confined to those that are undesirable. (For example, murder is subject to criminal sanctions, but not all accidental killing is.) As we have stressed, when the socially costly sanction of imprisonment is employed, the fault system is desirable because it results in less frequent imposition of punishment than strict liability. Also, the focus on intent in criminal law as a precondition for imposing sanctions may be sensible with regard to deterrence because those who intend to do harm are more likely to conceal their acts, and may be harder to discourage because of the benefits they anticipate. That unsuccessful

attempts to do harm are punished in criminal law is an implicit way of raising the likelihood of sanctions for undesirable acts. Study of specific doctrines of criminal law seems to afford a rich opportunity for economic analysis.

Criticism of Economic Analysis of Law

Many observers, and particularly non-economists, view economic analysis of law with scepticism. We consider several such criticisms here.

Description of Behaviour

It is sometimes claimed that individuals and firms do not respond to legal rules as rational maximizers of their well-being. For example, it is often asserted that decisions to commit crimes are not governed by economists' usual assumptions. Some sceptics also suggest that, in predicting individuals' behaviour, certain standard assumptions are inapplicable. For example, in predicting compliance with a law, the assumption that preferences be taken as given would be inappropriate if a legal rule would change people's preferences, as some say was the case with civil rights laws and environmental laws. In addition, laws may frame individuals' understanding of problems, which could affect their probability assessments or willingness to pay. The emerging field of behavioural economics, as well as work in various disciplines that address social norms, is beginning to examine these sorts of issues (Jolls et al. 1998).

Distribution of Income

A frequent criticism of economic analysis of law concerns its focus on efficiency to the exclusion of the distribution of income. The claim of critics is that legal rules should be selected in a manner that takes into account their effects on the rich and the poor. But achieving sought-after redistribution through income tax and transfer programmes tends to be superior to redistribution through the choice of legal rules. This is because redistribution through legal rules and the tax-transfer system both will distort individuals' labour-leisure decisions in the same manner, but redistribution through legal rules often will require choosing an inefficient rule, which imposes an additional cost (Shavell 1981; Kaplow and Shavell 1994b).

Moreover, it is difficult to redistribute income systematically through the choice of legal rules. Many individuals are never involved in litigation; and for those who are there is substantial income heterogeneity among plaintiffs as well as among defendants. Additionally, in contractual contexts the choice of a legal rule often will not have any distributional effect because contract terms, notably the price, will adjust, so that any agreement into which parties enter will continue to reflect the initial distribution of bargaining power between them.

Concerns for Fairness

An additional criticism is that the conventional economic approach slights important concerns about fairness, justice and rights. Some of these notions refer implicitly to the appropriateness of the distribution of income and, accordingly, are encompassed by our preceding remarks. Also, to some degree, the notions are motivated by instrumental concerns. For example, the attraction of paying fair compensation to victims must derive in part from the beneficial risk reduction effected by such payments, and the appeal of obeying contractual promises must rest in part on the desirable consequences contract performance has on production and exchange. To some extent, therefore, critics' concerns are already taken into account in standard economic analysis.

However, many who promote fairness, justice and rights do not regard these notions merely as some sort of proxy for attaining instrumental objectives. Instead, they believe that satisfying these notions is intrinsically valuable. This view also can be partially reconciled with the economic conception of social welfare: if individuals have a preference for a legal rule or institution because they regard it as fair, that should be credited in the determination of social welfare, just as any other preference should.

But many commentators take the position that conceptions of fairness are important as ethical principles in themselves, without regard to any possible relationship the principles may have to individuals' welfare. This opinion is the subject of

long-standing debate among moral philosophers. Some readers may be sceptical of normative views that are not grounded in individuals' well-being because embracing such views entails a willingness to sacrifice individuals' well-being. Indeed, consistently pursuing any non-welfarist principle must sometimes result in everyone being made worse off (see Kaplow and Shavell 2001, 2002).

Efficiency of Judge-Made Law

Also criticized is the contention of some economically oriented legal academics, notably Posner (1972), that judge-made law tends to be efficient (in contrast to legislation, which is said to reflect the influence of special interest groups). Some critics believe that judge-made law is guided by notions of fairness, or is influenced by legal culture or judges' biases, and thus will not necessarily be efficient. Whatever is the merit of the critics' claims, they are descriptive assertions about the law, and their validity does not bear on the power of economics to predict behaviour in response to legal rules or on the value of normative economic analysis of law.

See Also

- ▶ Coase Theorem
- ▶ Law, Public Enforcement of
- ▶ Property Law, Economics and
- ▶ Uncertainty
- ▶ Welfare Economics

Bibliography

Ayres, I., and R. Gertner. 1989. Filling gaps in incomplete contracts: An economic theory of default rules. *Yale Law Journal* 99: 87–130.

Bebchuk, L. 1984. Litigation and settlement under imperfect information. *RAND Journal of Economics* 15: 404–415.

Bebchuk, L., and O. Ben-Shahar. 2001. Pre-contractual reliance. *Journal of Legal Studies* 30: 423–457.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Becker, G., and G. Stigler. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Bentham, J. 1789. *An introduction to the principles of morals and legislation, in the utilitarians*. Garden City: Anchor Books, 1973.

Blume, L., D. Rubinfeld, and P. Shapiro. 1984. The taking of land: When should compensation be paid? *Quarterly Journal of Economics* 99: 71–92.

Bolton, P., and M. Dewatripont. 2005. *Contract theory*. Cambridge, MA: MIT Press.

Brown, J. 1973. Toward an economic theory of liability. *Journal of Legal Studies* 2: 323–349.

Calabresi, G. 1961. Some thoughts on risk distribution and the law of torts. *Yale Law Journal* 70: 499–553.

Calabresi, G. 1970. *The costs of accidents: A legal and economic analysis*. New Haven: Yale University Press.

Calabresi, G., and A. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review* 85: 1089–1128.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Cooter, R., and T. Ulen. 2003. *Law and economics*. 4th ed. Reading, MA: Addison-Wesley.

Craswell, R. 1996. Offer, acceptance, and efficient reliance. *Stanford Law Review* 48: 481–553.

Daughety, A. 2000. Settlement. In *Encyclopedia of law and economics*, ed. B. Bouckaert and G. De Geest, Vol. 5. Cheltenham: Edward Elgar.

Daughety, A., and J. Reinganum. 2000. Appealing judgments. *RAND Journal of Economics* 31: 502–525.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review: Papers and Proceedings* 57: 347–359.

Diamond, P. 1974. Single activity accidents. *Journal of Legal Studies* 3: 107–164.

Diamond, P., and E. Maskin. 1979. An equilibrium analysis of search and breach of contract, I: Steady states. *Bell Journal of Economics* 10: 282–316.

Hadfield, G. 1994. Judicial competence and the interpretation of incomplete contracts. *Journal of Legal Studies* 23: 159–184.

Hart, O. 1987. Incomplete contracts. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, Vol. 2. New York: Macmillan.

Hart, O., and J. Moore. 1988. Incomplete contracts and renegotiation. *Econometrica* 56: 755–758.

Hay, B., and K. Spier. 1998. Settlement of litigation. In *The new palgrave dictionary of economics and the law*, ed. P. Newman, Vol. 3. London: Macmillan.

Innes, R. 1999. Remediation and self-reporting in optimal law enforcement. *Journal of Public Economics* 72: 379–393.

Jolls, C., C. Sunstein, and R. Thaler. 1998. A behavioral approach to law and economics. *Stanford Law Review* 50: 1471–1550.

Kaplow, L. 1986. An economic analysis of legal transitions. *Harvard Law Review* 99: 509–617.

Kaplow, L. 1994. The value of accuracy in adjudication: An economic analysis. *Journal of Legal Studies* 23: 307–401.

Kaplow, L., and S. Shavell. 1994a. Optimal law enforcement with self-reporting of behavior. *Journal of Political Economy* 102: 583–606.

Kaplow, L., and S. Shavell. 1994b. Why the legal system is less efficient than the income tax in redistributing income. *Journal of Legal Studies* 23: 667–681.

Kaplow, L., and S. Shavell. 2001. Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy* 109: 281–286.

Kaplow, L., and S. Shavell. 2002. *Fairness versus welfare*. Cambridge, MA: Harvard University Press.

Katz, A. 1990. The strategic structure of offer and acceptance: Game theory and the law of contract formation. *Michigan Law Review* 89: 215–295.

Katz, A. 1996. When should an offer stick? The economics of promissory estoppel in preliminary negotiations. *Yale Law Journal* 105: 1249–1309.

Kronman, A. 1978. Mistake, disclosure, information, and the law of contracts. *Journal of Legal Studies* 7: 1–34.

Landes, W., and R. Posner. 1975. The private enforcement of law. *Journal of Legal Studies* 4: 1–46.

Landes, W., and R. Posner. 1987a. *The economic structure of tort law*. Cambridge, MA: Harvard University Press.

Landes, W., and R. Posner. 1987b. Trademark law: An economic perspective. *Journal of Law and Economics* 30: 265–309.

Landes, W., and R. Posner. 2003. *The economic structure of intellectual property law*. Cambridge, MA: Harvard University Press.

Libecap, G. 1986. Property rights in economic history: Implications for research. *Explorations in Economic History* 23: 227–252.

Miceli, T. 1997. *Economics of the law: Torts, contracts, property, litigation*. New York: Oxford University Press.

Mookherjee, D., and I. Png. 1992. Monitoring vis-à-vis investigation in enforcement of law. *American Economic Review* 82: 556–565.

Png, I. 1986. Optimal subsidies and damages in the presence of judicial error. *International Review of Law and Economics* 6: 101–105.

Polinsky, A.M. 1980. Private versus public enforcement of fines. *Journal of Legal Studies* 9: 105–127.

Polinsky, A.M. 1983. Risk sharing through breach of contract remedies. *Journal of Legal Studies* 12: 427–444.

Polinsky, A.M., and S. Shavell. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69: 880–891.

Polinsky, A.M., and S. Shavell. 2000. The economic theory of public enforcement of law. *Journal of Economic Literature* 38: 45–76.

Polinsky, A.M., and S. Shavell. 2001. Corruption and optimal law enforcement. *Journal of Public Economics* 81: 1–24.

Polinsky, A.M., and S. Shavell, ed. 2007. *Handbook of law and economics*. Vol. 1. Amsterdam: North-Holland.

Posner, R. 1972. *Economic analysis of law*. Boston: Little, Brown and Company.

Posner, R. 1985. An economic theory of the criminal law. *Columbia Law Review* 85: 1193–1231.

Posner, R. 2003. *Economic analysis of law*. 6th ed. New York: Aspen Publishers.

Priest, G., and B. Klein. 1984. The selection of disputes for litigation. *Journal of Legal Studies* 13: 1–55.

Reinganum, J. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78: 713–728.

Reinganum, J., and L. Wilde. 1986. Settlement, litigation, and the allocation of litigation costs. *RAND Journal of Economics* 17: 557–566.

Rogerson, W. 1984. Efficient reliance and damage measures for breach of contract. *RAND Journal of Economics* 15: 39–53.

Rose-Ackerman, S. 1999. *Corruption and government: Causes, consequences and reform*. New York: Cambridge University Press.

Schwartz, A. 1992. Relational contracts in the courts: An analysis of incomplete agreements and judicial strategies. *Journal of Legal Studies* 21: 271–318.

Schweizer, U. 1989. Litigation and settlement under two-sided incomplete information. *Review of Economic Studies* 56: 163–178.

Shavell, S. 1980a. Damage measures for breach of contract. *Bell Journal of Economics* 11: 466–490.

Shavell, S. 1980b. Strict liability versus negligence. *Journal of Legal Studies* 9: 1–25.

Shavell, S. 1981. A note on efficiency vs distributional equity in legal rulemaking: Should distributional equity matter given optimal income taxation? *American Economic Review: Papers and Proceedings* 71: 414–418.

Shavell, S. 1982a. On liability and insurance. *Bell Journal of Economics* 13: 120–132.

Shavell, S. 1982b. The social versus the private incentive to bring suit in a costly legal system. *Journal of Legal Studies* 11: 333–339.

Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85: 1232–1262.

Shavell, S. 1987a. *Economic analysis of accident law*. Cambridge, MA: Harvard University Press.

Shavell, S. 1987b. The optimal use of nonmonetary sanctions as a deterrent. *American Economic Review* 77: 584–592.

Shavell, S. 1987c. A model of optimal incapacitation. *American Economic Review: Papers and Proceedings* 77: 107–110.

Shavell, S. 1989. Sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20: 183–195.

Shavell, S. 1991. Specific versus general enforcement of law. *Journal of Political Economy* 99: 1088–1108.

Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36: 255–287.

Shavell, S. 1994. Acquisition and disclosure of information prior to sale. *RAND Journal of Economics* 25: 20–36.

Shavell, S. 1995. The appeals process as a means of error correction. *Journal of Legal Studies* 24: 379–426.

Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26: 575–612.

Shavell, S. 2004. *Foundations of economic analysis of law*. Cambridge, MA: Harvard University Press.

Shavell, S. 2006. On the writing and interpretation of contracts. *Journal of Law, Economics, & Organization* 22: 289–314.

Shleifer, A., and R. Vishny. 1993. Corruption. *Quarterly Journal of Economics* 108: 599–617.
Spier, K. 1992. The dynamics of pretrial negotiation. *Review of Economic Studies* 59: 93–108.
Spitzer, M., and E. Talley. 2000. Judicial auditing. *Journal of Legal Studies* 29: 649–683.
Umbeck, J. 1981. *A theory of property rights with application to the California gold rush*. Ames: Iowa State University Press.
Wils, W. 1993. Who should bear the costs of failed negotiations? A functional inquiry into precontractual liability. *Journal des Economistes et des Etudes Humaines* 4: 93–134.

Law, John (1671–1729)

Michael D. Bordo

Keywords

Banque Royale; Commodity money; Convertibility; Disequilibrium theory of money; Fiat money; Law, J.; Medium of exchange; Mississippi Bubble; Money supply; National banks; Paper money; Real bills doctrine; Speculation

JEL Classifications

B31

John Law of Lauriston has been regarded by some observers as a monetary crank, by others as a precursor of modern schemes of managed money and Keynesian full – employment policies. He was the originator of the Mississippi Bubble, perhaps the greatest speculative bubble of all time.

Born in Edinburgh, the son of prosperous parents, Law was well educated in political economy. A fugitive from justice in 1694 for killing a man in a duel in England, Law travelled extensively throughout Europe, observing and gaining experience in banking, insurance and finance. He proposed a number of unsuccessful schemes to set up a national bank of issue – in Paris in 1702, Edinburgh in 1705 and Savoy in 1712 – finally attaining success in France with the establishment in 1718 of the Banque Royale.

Law's theories on money and banking are contained in *Money and Trade Considered: With a Proposal for Supplying the Nation With Money* (1705) and other works (Hamilton 1968; Harsin 1934). Like other 18th-century writers Law adopted a disequilibrium theory of money, viewing it as a stimulant to trade. In a state of unemployment, Law maintained that an increase in the nation's money supply would stimulate employment and output without raising prices since the demand for money would rise with the increase in output. Moreover, once full employment was attained the monetary expansion would attract factors of production from abroad, so output would continue to increase.

According to Law, a paper-money standard was preferable to one based on precious metals. Suitable candidates for the money supply included government fiat, banknotes, stocks and bonds. Since the primary function of money was as a medium of exchange, it could best be served by a commodity (paper) not subject to considerable fluctuation in value and high resource costs. Thus Law advocated the establishment of note-issuing national banks that would extend productive loans (real bills), providing sufficient currency to guarantee prosperity. Two proposals for such banks, in Paris 1702 and Edinburgh 1705, would have had the note issues based on land initially valued in terms of silver.

From 1716 to 1720 John Law had the unique opportunity to apply his theories to the French economy. In 1715, the heritage of two exhausting wars was depression and deflation. Law succeeded in convincing the Regent (the Duke of Orleans) that a bank of issue would alleviate the problem of financing the national debt.

Accordingly, he established in Paris on 2 May 1716 a private bank, the Banque Générale. In its 31 months of operation, the bank was remarkably successful; its notes (convertible into specie and payable as taxes) were issued in moderation and gained national circulation. On 4 December 1718, the Banque Générale was nationalized and renamed the Banque Royale, with Law in control, and in January 1719 it began to issue notes denominated in *livres tournois,* the unit of account, replacing the previously issued *écus de banque* representing fixed amounts of specie.

Alongside the bank, in August 1717, Law established the Compagnie d'Occident after obtaining the franchise on Louisiana and the monopoly of the Canadian fur trade. This company in the succeeding 22 months acquired the tobacco monopoly, the East India Company and the trading monopolies to Africa and China. Law changed its name in June 1719 to the Compagnie des Indes, and the following winter obtained the farm of the royal mints and of the indirect taxes. In October 1719 he refunded the national debt of 1.5 million *livres tournois,* and in January 1720 became Finance Minister.

The stock of the Compagnie des Indes, initially selling at a par value of £500, within half a year in an unprecedented speculative mania was bid up to many times its original price. The bubble burst in January 1720 after the price of the stock reached a peak of £18,000. To support the price Law made the mistake of pegging it at £9,000, thereby monetizing it and engendering a rapid expansion of notes (125per cent in two months). In May 1720, in a desperate attempt to salvage his system Law issued a deflationary decree depreciating the stock and reducing the denomination of notes by stages. This decree led to a panic as the public, fearful of further capital losses, sold off both notes and stock. Law's dismissal by the Regent worsened the panic. He was quickly reinstated but his final attempt to restore confidence by reducing the outstanding note issue proved unsuccessful. By December 1720 the 'system' collapsed. Law fled to Belgium and payments quickly reverted to a specie basis. The collapse of the system ruined many in all walks of life and made the word 'bank' anathema in France for well over a century.

Though Law's system reduced unemployment and stimulated output, it was at the expense of doubling the price level. His system was undermined by his actions breaking the link between the note issue and specie convertibility; by retiring the national debt with bank notes convertible into stock; and by encouraging speculation in stock by declaring dividends unrelated to the company's true prospects.

Monetizing the stock by pegging its price in the end destroyed the public's confidence in his system. Law was aware of many of the principles of sound money and banking, but by equating money with stock and relying on the real bills doctrine he sowed the seeds of disaster.

Selected Works

1705. *Money and trade considered: With a proposal for supplying the nation with money.* New York: Augustus Kelley, 1966.

Bibliography

Blaug, M. 1978. *Economic theory in retrospect,* 3rd ed. Cambridge: Cambridge University Press.

Hamilton, E.J. 1936. Prices and wages at Paris under John Law's system. *Quarterly Journal of Economics* 51: 42–70.

Hamilton, E.J. 1968. Law, John. In *International encyclopedia of the social sciences*, vol. 9. New York: The Free Press.

Hamilton, E.J. 1969. The political economy of France at the time of John Law. *History of Political Economy* 1: 123–149.

Harsin, P. 1934. *John Law: Oeuvres complètes,* 3 vols. Paris: Sirey.

Kindleberger, C.P. 1984. *A financial history of Western Europe*. London: Allen & Unwin.

Neal, L., and E. Schubert. 1985. The first rational bubble: A new look at the Mississippi and South Sea schemes. Faculty Working Paper No. 1188, Bureau of Economic and Business Research, University of Illinois at Urbana-Champaign.

Rist, C.W. 1940. *A history of monetary and credit theory from John Law to the present day.* London: Allen & Unwin.

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

Law, Public Enforcement of

Mitchell Polinsky and Steven Shavell

Abstract

This article surveys the economic analysis of public enforcement of law – the use of public agents (inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of

legal rules. We first discuss the basic elements of the theory: the probability of imposition of sanctions, the magnitude and form of sanctions (fines, imprisonment), and the rule of liability. We then examine a variety of extensions, including the costs of imposing fines, mistakes, marginal deterrence, settlement, self-reporting, repeat offences, and incapacitation.

Keywords

Audits; Becker, G; Bentham, J; Bribery; Corruption; Criminal law; Deterrence; Efficiency wages; Fairness; Fault-based liability; Fines; Harm; Imprisonment; Incapacitation; Principal and agent; Public enforcement of law; Safety regulations; Sanctions; Self-reporting; Settlements; Social norms; Strict liability; Time discount rates

JEL Classification

H23; H26; J28; K14; K32; K42

In this article we consider the theory of public enforcement of law – the use of public agents (inspectors, tax auditors, police, prosecutors) to detect and to sanction violators of legal rules. After briefly discussing the rationale for public (as opposed to private) enforcement, we present the basic elements of the theory: the probability of imposition of sanctions, the magnitude and form of sanctions (fines, imprisonment), and the rule of liability. We then examine a variety of extensions of the central theory, including the costs of imposing fines, mistakes, marginal deterrence, settlement, self-reporting, repeat offences, and incapacitation. (For a fuller treatment of the material in this entry, see Polinsky and Shavell 2007.)

Before proceeding, we note that economically oriented analysis of public law enforcement dates primarily from the eighteenth century contribution of Jeremy Bentham (1789), whose analysis of deterrence was sophisticated and expansive. After Bentham, the subject of enforcement lay essentially dormant in economic scholarship until Gary Becker (1968) published a highly influential article, which has led to a voluminous literature.

Rationale of Public Enforcement

A basic question is why there is a need for public enforcement of law (see generally Becker and Stigler 1974; Landes and Posner 1975; Polinsky 1980a). In particular, why not rely solely on private suits brought by victims? The answer depends importantly on the locus of information about the identity of injurers. When victims of harm naturally possess knowledge of the identity of injurers, allowing private suits for damages will motivate victims to sue and thus harness the information they have for purposes of law enforcement. This may explain why the enforcement of contractual obligations and of accident law is primarily private. When victims do not know who caused harm, however, or when finding injurers is difficult, society may need to rely instead on public investigation and prosecution; this is broadly true of crimes and of many violations of environmental and safety regulations.

Basic Framework for Analysing Public Enforcement

An individual who commits a harmful act obtains a gain and also faces the risk of being caught and sanctioned. The form of sanction could be a fine or a prison term. Fines generally will be treated as socially costless because they are mere transfers of money, whereas imprisonment will be considered as socially costly because of the expense of operating prisons and the disutility suffered by those imprisoned. The higher the probability of detecting violators, the more resources the state must devote to enforcement.

We assume that social welfare equals the sum of individuals' expected utilities. If individuals are risk neutral, social welfare can be expressed as the gains individuals obtain from committing their harmful acts, minus the harms caused and the costs of law enforcement. The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures (or, equivalently, a probability of detection), the form of sanctions, and their level.

Fines

Suppose that the sanction is a fine and that individuals are risk neutral. If the probability of detection p is taken as fixed, then the optimal fine is the harm h divided by the probability, that is, h/p; for then the expected fine $p(h/p)$ equals h. This fine is optimal because, facing it, an individual will commit a harmful act if, and only if, the gain he would derive exceeds the harm he would cause. Such behaviour is first-best. The fundamental formula h/p essentially was noted by Bentham (1789) and it has been observed by many others since.

If the probability of detection can be varied, the optimal fine is maximal, f_M, as emphasized by Becker (1968). If the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence. If $f < f_M$, then raise the fine to f_M and lower the probability from p to $(f/f_M)p$; the expected fine is still pf, so that deterrence is maintained but expenditures on enforcement are reduced, implying that social welfare rises.

The optimal probability p of imposing a fine is low in the sense that it results in some under-deterrence; that is, the optimal p is such that the expected fine pf_M is less than the harm h (Polinsky and Shavell 1984). The reason is to economize on enforcement resources. In particular, if pf_M equals h, behaviour will be ideal, meaning that the individuals who are just deterred obtain gains essentially equal to the harm. These are the individuals who would be led to commit the harmful act if p were lowered slightly. That in turn must be socially beneficial because these individuals cause no net social losses (their gains essentially equal the harm), but reducing p saves enforcement costs. How much pf_M should be lowered below h depends on the saving in enforcement costs from reducing p compared with the net social costs of under-deterrence that will result if p is lowered non-trivially.

If individuals are risk averse, the optimal fine may be well less than the maximal fine, as first shown in Polinsky and Shavell (1979); see also Kaplow (1992). This is because a high fine would impose substantial risk-bearing costs on individuals who commit harmful acts. If $f < f_M$, it is still true that f can be raised and p lowered so as to maintain deterrence, but because of risk aversion this now implies that pf falls, meaning that fine revenue falls. The reduction in fine revenue reflects the disutility caused by imposing greater risk on risk-averse individuals. The decline in fine revenue could more than offset the savings in enforcement expenditures, causing social welfare to be lower.

Imprisonment

Now suppose that the sanction is imprisonment. If the probability of detection is fixed, there is no simple formula for the optimal imprisonment term (see Polinsky and Shavell 1984). The optimal term could be such that there is either under-deterrence or over-deterrence. On the one hand, a relatively low imprisonment term, implying under-deterrence, might be socially desirable because imprisonment costs are reduced for those individuals who commit harmful acts. On the other hand, a relatively high term, implying over-deterrence, might be socially desirable because imprisonment costs are reduced due to fewer individuals committing harmful acts, even if some of these deterred individuals would have obtained gains exceeding the harm.

If the probability of detection can be varied and individuals are risk neutral in imprisonment, then the optimal imprisonment term is maximal. The reasoning is similar to that employed above: if the imprisonment term were not maximal, it could be raised and the probability of detection lowered so as to keep the expected prison term constant; neither individual behaviour nor the costs of imprisonment are affected, but enforcement expenditures fall.

If, instead, individuals are risk averse in imprisonment (the disutility of each additional year of imprisonment grows with the number of years in prison), there is a stronger argument for setting the imprisonment sanction maximally (Polinsky and Shavell 1999). Now when the imprisonment term is raised, the probability of detection can be lowered more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but also the costs of imposing

imprisonment sanctions decline because the expected prison term falls.

Last, suppose that individuals are risk preferring in imprisonment (the disutility of each additional year of imprisonment declines with the number of years in prison). This possibility seems particularly important: the first years of imprisonment may create unusually high disutility, due to brutalization of the prisoner or to the stigma of having been imprisoned at all. Individuals' positive time discount rates, which are thought to be especially significant for criminals, also make the disutility of later years less significant. In the case of risk-preferring individuals, the optimal prison term may well be less than maximal: if the sentence were raised, the probability that maintains deterrence could not be lowered proportionally, implying that the expected prison term would rise. Thus, although there would be enforcement-cost savings, they might not be great enough to offset the increased sanctioning costs.

When the sanction is imprisonment, the optimal probability of detection may be such that there is either under-deterrence or over-deterrence. On the one hand, the motive to lower the probability is reinforced relative to the case of fines because imprisonment costs, as well as detection costs, decline if fewer offenders are caught. On the other hand, raising the probability of detection results in fewer offenders, which, everything else equal, decreases imprisonment costs because fewer are imprisoned. Either effect may dominate.

Fines Versus Imprisonment

Fines generally are preferable to prison terms as a means of deterrence, since fines are socially cheaper sanctions to impose (Becker 1968; Polinsky and Shavell 1984). Hence, fines should be employed to the greatest extent possible – until a party's wealth is exhausted – before imprisonment is imposed. Further, imprisonment should be used as a sanction only if the harm prevented by the added deterrence is sufficiently great.

Fault-Based Liability

Our discussion thus far has presumed that liability is strict (imposed whenever harm occurs), but liability may instead be based on fault (imposed only when behaviour was found to be socially undesirable). Fault-based liability, like strict liability, can induce individuals to behave properly, but fault-based liability possesses an advantage when individuals are risk averse: if they act responsibly, they will not be found at fault, so will not bear the risk of being sanctioned. Similarly, fault-based liability is advantageous when the sanction is imprisonment, for then again individuals may be led to behave optimally without the actual imposition of sanctions, and thus without social costs being incurred (Shavell 1987b). To the extent that mistakes are made in determining fault, however, these two advantages are reduced.

Fault-based liability is more difficult to implement because it requires more information than strict liability. To apply fault-based liability, the enforcement authority must be able to determine the proper fault standard – that is, socially desirable behaviour – and it must ascertain whether the defendant's conduct was in compliance with the fault standard. Under strict liability, the authority need only measure harm. (Moreover, for reasons we discuss below, strict liability encourages better decisions by injurers regarding their level of participation in harm-creating activities.)

This concludes the presentation of the basic theory of public enforcement of law. We now turn to various extensions and refinements of the analysis.

Accidental Harms

We have been implicitly assuming that individuals decide whether or not to commit acts that cause harm with certainty, that is, they decide whether or not to cause intentional harms. In many circumstances, however, harms are accidental – they occur only with a probability. Essentially all that we have said above applies in a straightforward way when harms are accidental.

There is, however, an additional issue that arises when harm is uncertain: a sanction can be imposed either on the basis of the commission of an act that increases the chance of harm (such as

storing chemicals in a substandard tank) or on the basis of the actual occurrence of harm (if the tank ruptures and results in a spill). In principle, either approach can achieve optimal deterrence – by setting the (expected) sanction equal to expected harm if liability is imposed whenever a dangerous act is committed, or equal to actual harm if liability is imposed only if harm occurs.

Several factors are relevant to the choice between act-based and harm-based sanctions (Shavell 1993). First, act-based sanctions need not be as high as harm-based sanctions to accomplish a given level of deterrence (expected harm is less than actual harm), and thus offer an advantage because of parties' limited assets. Second, because act-based sanctions can accomplish a given level of deterrence with lower sanctions, they are preferable when parties are risk averse. Third, either act-based sanctions may be simpler to impose (it might be less difficult to determine whether an oil shipper properly maintains its vessels' holding tanks than to detect whether one of the vessels leaked oil), or harm-based sanctions may be easier to implement (a driver who causes harm might be caught without difficulty, but not one who speeds). Fourth, it may be hard to calculate the expected harm due to an act, but relatively easy to ascertain the actual harm if it eventuates, favoring harm-based sanctions.

Costs of Imposing Fines

The costs borne by enforcement authorities in imposing fines should be reflected in the fine. Recall that, if the probability of detection is taken as fixed and individuals are risk neutral, the optimal fine is h/p, the harm divided by the probability of detection. Now suppose there is a public cost k of imposing a fine. The optimal fine then becomes $h/p + k$; the cost k should be added to the fine that would otherwise be desirable (Becker 1968; Polinsky and Shavell 1992). The explanation is that, if an individual commits a harmful act, he causes society to bear not only the immediate harm h but also, with probability p, the cost k of imposing the fine – that is, his act results in an expected total social cost of $h + pk$. If the fine is $h/p + k$, the individual's expected fine is $p[h/p + k] = h + pk$, leading him to commit the harmful act if and only if his gain exceeds the expected total social cost of his act.

Not only does the state bear costs when fines are imposed, so do individuals who pay the fines (such as legal defence expenses). The costs borne by individuals, however, do not affect the formula for the optimal fine. Individuals properly take these costs into account because they bear them.

Level of Activity

In many settings in which harm may occur, an individual chooses not only whether to commit a harmful act when engaging in an activity, but also the level at which to engage in the activity. Drivers decide how careful to be while driving, as well as how many miles to drive; similarly, firms choose safety precautions as well as their level of output. The socially optimal activity level is such that the actor's marginal utility from the activity just equals the marginal expected harm caused by the activity (we assume that optimal care is taken). Thus, the optimal number of miles driven is the level at which the marginal utility of driving an extra mile just equals the marginal expected harm per mile driven.

Under strict liability parties will choose the optimal level of activity because they will pay for all harm done. They will choose the optimal number of miles to drive because they will pay for all harm per mile driven. Under fault-based liability, however, parties generally do not pay for the harm they cause because they tend to behave so as not to be found at fault. As a consequence, they will choose an excessive level of activity (Shavell 1980). Driving more miles increases expected harm, but this effect generally will be ignored under fault-based liability.

The interpretation of the preceding points in relation to firms is that under strict liability the product price will reflect the expected harm caused by production. Hence, the amount purchased, and thus the level of production, will tend to be socially optimal. However, under fault-based liability the product price will not

reflect harm, but only the cost of precautions; thus, the level of output will be excessive (Polinsky 1980b).

Relatedly, safety regulations and other regulatory requirements are often framed as standards of care that have to be met, but which, if met, free the regulated party from liability. Hence, regulations of this sort are subject to the criticism that they lead to excessive levels of the regulated activity. Making parties strictly liable for harm would be superior to safety regulation with respect to inducing socially correct activity levels.

Mistakes

An individual who should be found liable might mistakenly be acquitted. Conversely, an individual who should not be found liable might mistakenly be convicted. For an individual who has been detected, let the probabilities of these errors be ε_A and ε_C, respectively. Given the probability of detection p and the chances of these types of error, an individual will commit the wrongful act if and only if his gain g net of his expected fine if he does commit it exceeds his expected fine if he does not commit it, that is, when $g - p(1-\varepsilon_A)f > -p\varepsilon_C f$, or, equivalently, when $g > (1-\varepsilon_A-\varepsilon_C)pf$.

As emphasized by Png (1986), both types of error reduce deterrence: the term $(1-\varepsilon_A-\varepsilon_C)pf$ is declining in both ε_A and ε_C. The first type of error diminishes deterrence because it lowers the expected fine if an individual violates the law. The second type of error lowers deterrence because it reduces the difference between the expected fine from violating the law and not violating it—the greater is ε_C, the smaller is the increase in the expected fine if one violates the law.

Because mistakes dilute deterrence, they reduce social welfare. Specifically, to achieve any level of deterrence, the probability p must be higher to offset the effect of errors. Mistaken convictions have the additional effect of discouraging socially desirable participation in the activity. Consequently, expenditures made to reduce errors may be socially beneficial (Kaplow and Shavell 1994a).

Two other points regarding the implications of mistake are worth noting. First, if individuals are risk averse, the possibility of mistakes of either type generally lowers optimal sanctions (Block and Sidak 1980). Second, as stressed by Craswell and Calfee (1986), individuals will often have a motive to take excessive precautions under fault-based liability in order to reduce the chance of being found erroneously at fault.

General Enforcement

In many settings, enforcement may be said to be general in the sense that several different types of violations will be detected by an enforcement agent's activity. For example, a police officer waiting at the roadside may notice a driver who litters as well as one who goes through a red light or who speeds, and a tax auditor may detect a variety of infractions when he examines a tax return. (In contrast, if enforcement is specific, the probability is chosen independently for each type of harmful act.)

When enforcement is general, the optimal sanction rises with the level of harm, and is maximal only for relatively high harms (Shavell 1991; Mookherjee and Png 1992). To see why, assume that liability is strict, the sanction is a fine, and injurers are risk neutral. Let $f(h)$ be the fine given harm h. Then, for any general probability of detection p (that is, p applies regardless of h), the optimal fine schedule is h/p, provided that h/p is feasible; otherwise the optimal fine is maximal. This schedule is obviously optimal given p because it implies that the expected fine equals harm, thereby inducing ideal behaviour whenever that is possible. That sanctions should rise with the severity of harm up to a maximum when enforcement is general also holds if the sanction is imprisonment and if liability is fault-based.

Marginal Deterrence

In many circumstances a person may consider which of several harmful acts to commit: for example, whether to release only a small amount

of a pollutant into a river or a large amount, or whether to kidnap a person or also to kill the kidnap victim. In such contexts, sanctions influence which harmful acts individuals choose to commit (as well as whether to commit any harmful act). Marginal deterrence is said to occur when a more harmful act is deterred because its sanction exceeds that for a less harmful act (Stigler 1970; Shavell 1992; Wilde 1992; Mookherjee and Png 1994).

Other things being equal, it is socially desirable that enforcement policy creates marginal deterrence so that, when harmful acts do occur, less harm is done. One way to accomplish marginal deterrence is for sanctions to rise with the magnitude of harm, which means that sanctions generally will not be maximal. However, fostering marginal deterrence may conflict with achieving overall deterrence: in order for the schedule of sanctions to rise steeply enough to accomplish marginal deterrence, sanctions for less harmful acts may have to be so low that individuals are not deterred from committing some harmful act.

Note that marginal deterrence also can be promoted by increasing the probability of detection. Kidnappers can be better deterred from killing their victims if more police resources are devoted to apprehending kidnappers who murder their victims than to those who do not.

Principal–Agent Relationship

Although we have assumed that an injurer is a single actor, injurers often are more appropriately characterized as collective entities, and specifically as a principal and the principal's agent. For example, the principal could be a firm and the agent an employee, or the principal could be a contractor and the agent a subcontractor.

When harm is caused by the behaviour of principals and agents, many of our prior conclusions carry over to the sanctioning of principals. Notably, if a risk-neutral principal faces an expected fine equal to harm done, he will behave socially optimally in controlling his agents, and in particular will contract with them and monitor them in ways that will give the agents appropriate incentives to reduce harm (Newman and Wright 1990; but see Arlen 1994).

An issue that arises when there are principals and agents concerns the allocation of financial sanctions between the two parties. It is apparent that the particular allocation of sanctions does not matter when the parties can reallocate the sanctions through their own contract. For example, if the agent finds that he faces a large fine but is more risk averse than the principal, the principal can assume it; conversely, if the fine is imposed on the principal, he will retain it and not impose an internal sanction on the agent. Thus, the post-contract sanctions that the agent bears are not affected by the particular division of sanctions initially selected by the enforcement authority.

The allocation of monetary sanctions between principals and agents would matter, however, if some allocations allow the pair to reduce their total burden. An important example is when a fine is imposed only on the agent and he is unable to pay it (Sykes 1981; Kornhauser 1982). Then, he and the principal (who often would have higher assets) would jointly escape part of the fine, diluting deterrence. The fine therefore should be imposed on the principal rather than on the agent (or at least the part of the fine that the agent cannot pay).

A closely related point is that the imposition of imprisonment sanctions on agents may be desirable when their assets are less than the harm that they can cause, even if the principal's assets are sufficient to pay the optimal fine (Polinsky and Shavell 1993). That an agent's assets are limited means that the principal may be unable to control him adequately through the use of contractually determined penalties, which can only be monetary. In such circumstances it may be socially valuable to use the threat of a jail sentence to better control agents' misconduct.

Settlements

It is common for lawbreakers to settle with public enforcement authorities prior to being found liable in a trial. (In the criminal context, the settlement usually takes the form of a plea bargain, an

agreement in which the injurer pleads guilty to a reduced charge.) Both parties might prefer an out-of-court settlement to avoid the cost of a trial and to eliminate the risks inherent in the trial outcome (Cooter and Rubinfeld 1989; on plea bargaining, see Reinganum 1988, and Miceli 1996).

These advantages suggest that settlement is socially valuable, but the effect of settlement on deterrence is a complicating factor. Specifically, settlements dilute deterrence: for if injurers desire to settle, it must be because the expected disutility of sanctions is lowered for them (Polinsky and Rubinfeld 1988). The state may be able to offset this effect by increasing the level of sanctions.

Settlements may have other socially undesirable consequences. First, they may result in sanctions that are not as well tailored to harmful acts as would be true of court-determined sanctions. For example, if injurers have private information about the harm that they have caused, settlements will tend to reflect the average harm caused, resulting in high-harm (low-harm) injurers being under-deterred (over-deterred), whereas trial outcomes may better approximate the actual harm. Second, settlements hinder the amplification and development of the law through the setting of precedents. Third, if the sanction is imprisonment and defendants are risk averse, settlements necessitate longer terms than the expected sentence at trial in order to maintain deterrence, and thus increase public expenditures. On the social welfare evaluation of settlement, see, for example, Shavell (1997) and Spier (1997).

Self-Reporting

We have assumed that individuals are subject to sanctions only if they are detected by an enforcement agent, but in fact parties sometimes disclose their own violations. For example, firms often report infractions of environmental and safety regulations, individuals usually notify police of their involvement in traffic accidents, and even criminals occasionally turn themselves in.

Self-reporting can be induced by lowering the sanction for individuals who disclose their own violations (Kaplow and Shavell 1994b). Moreover, the reward for self-reporting can be made small, so that deterrence is only negligibly reduced. For example, if a risk-neutral individual commits a violation and does not self-report, his expected fine is pf. If he self-reports, the fine can be set just below pf, say at $pf - \varepsilon$, where $\varepsilon > 0$ is small. Then the individual will want to self-report but the deterrent effect of the sanction will be essentially the same as if he did not self-report.

There are several social advantages of self-reporting. First, self-reporting reduces enforcement costs because the enforcement authority does not have to identify and prove who the violator was. Second, self-reporting reduces risk (a relatively high sanction imposed with a relatively low probability is replaced by a certain punishment), and thus is advantageous if injurers are risk averse. Third, self-reporting may allow harm to be mitigated (early notice of an oil spill may facilitate its containment).

Repeat Offenders

In practice, the law often sanctions repeat offenders more severely than first-time offenders. This policy cannot be socially advantageous if deterrence always induces first-best behaviour. For if the expected sanction for an offence equals its harm, then raising the sanction because an offender has a record of sanctions would overdeter him. Only if deterrence is inadequate is it possibly desirable to condition sanctions on offence history to increase deterrence. But, as we observed above, it usually will be worthwhile for the state to tolerate some under-deterrence in order to reduce enforcement expenses.

If there is under-deterrence, making sanctions depend on offence history may be beneficial. First, the use of offence history may create an additional incentive not to violate the law: if detection results not only in an immediate sanction but also in a higher sanction for any future violation, an individual will, everything else equal, be deterred to a greater extent (Polinsky and Shavell 1998). Second, making sanctions depend on offence history allows society to take

advantage of information about the dangerousness of individuals and the need to deter them: individuals with offence histories may be more likely than average to commit future violations, which might make it desirable to impose higher sanctions on them (Rubinstein 1979; Polinsky and Rubinfeld 1991). In addition, if repeat offenders have higher propensities to commit violations, they are more likely to be worth incapacitating by imprisonment (see below).

Imperfect Knowledge About the Probability and Magnitude of Sanctions

Individuals might not know the true probability of a sanction because the enforcement authority refrains from publishing information about the probability (perhaps hoping that individuals will believe it to be higher than it is in fact); or because the probability depends on factors that individuals do not fully understand; or because probabilities are difficult to assess. Also, individuals may have incomplete knowledge of the true magnitude of sanctions, particularly if the levels of sanctions are discretionary.

The implications of injurers' imperfect knowledge are straightforward. First, to predict how individuals behave, what is relevant, of course, is not the actual probability and magnitude of a sanction but the perceived levels or distributions of these variables. Second, to determine the optimal probability and magnitude of a sanction, account must be taken of the relationship between the actual and the perceived variables (Bebchuk and Kaplow 1992; Kaplow 1990). For example, if enforcement resources are increased in order to raise the probability of detection, there might be a delay before this increase is perceived by individuals, making such an investment less worthwhile.

Incapacitation

Society may reduce harm not only through deterrence, but also by imposing sanctions that remove parties from positions in which they are able to cause harm, that is, by incapacitating them. Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their driver's licences, businesses can lose their rights to operate in certain markets, and the like.

Suppose that the sole function of imprisonment is to incapacitate. Then it will be desirable to keep someone imprisoned as long as the reduction in criminal harm from incapacitating him exceeds the cost of imprisonment (Shavell 1987c). Although this condition could hold for a long period, it often will not because the proclivity to commit crimes appears to decline sharply with age.

As a matter of economic logic, the incapacitation rationale might imply that a person should be imprisoned even if he has not committed a crime, because the danger he poses to society makes incapacitating him worthwhile. In practice, however, the commission of a harmful act may be a good basis for predicting a person's future behaviour, in which case the incapacitation rationale would suggest imprisoning an individual only if he has committed such an act.

Two observations are worth noting about the relationship between the incapacitation goal and the deterrence goal. First, when enforcement is based on incapacitation, the optimal magnitude of the sanction is independent of the probability of apprehension, which contrasts with the case when enforcement is based on deterrence. Second, when enforcement is deterrence-oriented, the probability and magnitude of sanctions depend on the ability to deter, and if this ability is limited (as, for instance, with the insane), a low expected sanction may be optimal, whereas a high sanction still might be called for to incapacitate.

Corruption

One form of corruption in the enforcement process is bribery, in which an enforcer accepts a payment in return for not reporting a violation (or for reducing the mandated sanction for the violation). A second form of corruption is framing and framing-related extortion, in which an enforcement agent may frame an innocent individual or threaten to frame him in order to extort money from him. On corruption of law

enforcement, see Bowles and Garoupa (1997) and Polinsky and Shavell (2001) (and on corruption more generally, see, for example, Shleifer and Vishny 1993, and Rose-Ackerman 1999).

Bribery dilutes deterrence of violations of law because it results in a lower payment by an individual than the sanction for the offence. Framing and framing-related extortion also dilute deterrence. The reason is that framing and extortion imply that those who act innocently face an expected sanction, so that the difference between the expected sanction if an individual commits a violation and if he does not is lessened. (This point is essentially the same as the earlier observation that mistaken convictions dilute deterrence.)

One way to reduce corruption is to impose fines (or imprisonment sentences) on individuals caught engaging in bribery, extortion or framing. Corruption also can be reduced by paying enforcers rewards for reporting violations. Such payments will reduce their incentive to accept bribes because they will sacrifice their rewards if they fail to report violations. But high rewards give enforcers a greater incentive to frame innocent individuals. A third way to control corruption is to pay enforcers more than their reservation wage (that is, to pay them an efficiency wage). Then they would have more to lose if punished for corrupt behaviour and denied future employment.

A natural question is whether the deterrence-diluting effects of corruption can be offset by raising the fine on offenders. In the basic risk-neutral model of enforcement, it is not possible to raise the fine because the optimal fine is maximal. More realistically, however, the optimal fine is less than maximal for a variety of reasons, including those related to risk aversion, marginal deterrence, and general enforcement. While it would then be possible to raise the fine to offset the deterrence-diluting effects of corruption, doing so would lead to social costs (for example, by imposing greater risk).

Costly Observation of Wealth

Individuals and firms may be able to hide assets from government enforcers, including by hoarding cash, transferring assets to relatives or related legal entities, or moving money to offshore bank accounts. Consequently, an individual's level of wealth might not be able to be observed at all, or only after a costly audit.

Suppose first that the enforcement authority employs fines as sanctions and can audit an individual who claims that he cannot pay the fine (Polinsky 2006a, b). The optimal fine for misrepresenting one's wealth level equals the fine for the offence divided by the audit probability, and therefore generally exceeds the fine for the offence. This is a natural generalization of the formula for the optimal fine when the probability of detection is fixed, which is the harm divided by the probability. Auditing is valuable because it reduces misrepresentation of wealth and thereby increases deterrence.

Next, suppose that the enforcement authority cannot observe wealth because the cost of an audit is prohibitively high (Levitt 1997; Polinsky 2006a, b). If the authority would have used fines alone if it could have observed wealth at no cost, it would have imposed a higher fine on higher-wealth individuals. It obviously cannot do this when wealth is unobservable. Instead, it may be desirable to use the threat of an imprisonment sentence to induce individuals capable of paying a higher fine to do so. Alternatively, the enforcement authority might have used both fines and imprisonment if it could have observed wealth at no cost. Perhaps surprisingly, the inability to observe wealth might not be detrimental in this case. The reason is that the mix of fines and imprisonment that would be chosen when wealth is observable might impose a higher burden (though a lower fine) on low-wealth individuals. Then, high-wealth individuals will naturally want to identify themselves. Specifically, they will prefer to pay a higher fine and bear a shorter imprisonment sentence than to masquerade as low-wealth individuals, who will bear longer imprisonment sentences and a higher overall burden.

Social Norms

To some extent, social norms and morality are substitutes for public law enforcement because

they encourage in significant ways the attainment of desired behaviour (McAdams and Rasmusen 2007; Posner 1997; Shavell 2002). Social norms influence behaviour partly through internal incentives: when a person obeys a moral rule, he will tend to feel virtuous, and if he disobeys the rule, he will tend to feel guilty. Social norms also affect behaviour through external incentives: when a person is observed by another party to have obeyed a moral rule, that party may bestow praise on the first party, who will enjoy the praise; and if the person is observed by the other party to have disobeyed the rule, the second party will tend to disapprove of the first party, who will dislike the disapproval. Because social norms channel behaviour in this way, some socially desirable conduct can be encouraged reasonably well without employing the legal system.

Notwithstanding these observations, there will, of course, often be a need for formal law enforcement. First, much conduct that society desires cannot be controlled through moral incentives alone. One reason is that the private gains from undesirable conduct are often large and dominate the moral incentives. Another reason is that external moral sanctions might be imposed only with a low probability (the robber, tax cheat or polluter might not be spotted by others). A second rationale of formal law enforcement is that the social harm from failing to control an act through moral incentives may be large. This makes the expense of law enforcement worth incurring (as in the case of controlling robbery, but not of breaking into a queue at a movie theatre).

Fairness

So far we have not considered the possibility that individuals have opinions about the fairness of sanctions or the arbitrariness of enforcement (Polinsky and Shavell 2000b; Kaplow and Shavell 2002). Suppose, first, that individuals believe that the magnitude of sanctions should reflect the gravity of the acts. As discussed previously, if individuals are risk neutral, the usual solution to the enforcement problem consists of the highest possible sanction and a relatively low probability of detection. When the issue of fairness is added to the analysis, however, the usual solution generally is not optimal because a very high sanction will be seen as unfair.

A consequence of the desire to keep sanctions at fair levels, meaning at quite constrained levels for acts that are not very harmful, is that the socially optimal probability of detection changes. The optimal probability could be higher than the conventionally optimal probability: to achieve a desired level of deterrence with a lower fairness-restricted sanction, the probability has to rise, perhaps significantly. Alternatively, the optimal probability could be lower than in the conventional case: the additional deterrence from raising the probability might be relatively low because the sanction is relatively low; and the lower the deterrent benefit from raising the probability, the lower would be the social incentive to devote resources to enforcement.

Another aspect of fairness concerns the probability of detection rather than the magnitude of sanctions. Suppose that individuals consider it unfair for some lawbreakers to be sanctioned when others, who were lucky enough not to be caught, are not sanctioned. Then the optimal probability would be higher, and therefore the optimal sanction would be lower, than in the absence of this fairness concern.

A further notion of fairness involves the form of liability, whether liability is strict or based on fault. Individuals might prefer fault-based liability because sanctions are imposed on parties only if they behaved in a socially inappropriate way.

A final issue concerns the relevance of fairness considerations when firms, as opposed to individuals, are sanctioned. If what matters in terms of fairness is that the individuals responsible for harmful acts bear sanctions, as opposed to the artificial legal entity of a firm, one would want to identify the sanctions actually suffered by such persons within a firm if the firm bears a sanction. Note, too, that the imposition of sanctions on firms often penalizes individuals who are unlikely to be considered responsible for the harm, namely, shareholders and customers.

Criminal Law

The subject of criminal law may be viewed in the light of the theory of public law enforcement (Posner 1985; Shavell 1985). First, the fact that the acts in the core area of crime (robbery, murder, rape, and so forth) are punished by the sanction of imprisonment makes basic sense. Were society to rely on fines alone, deterrence of the acts in question would be grossly inadequate. This is because the probability of detecting many of these acts is low, making the money sanction necessary for deterrence high, but the assets of individuals who commit these acts often are insubstantial. Hence, the threat of prison is needed for deterrence. Moreover, the incapacitative aspect of imprisonment is valuable because of the difficulty of deterring individuals who are prone to commit criminal acts.

Second, many of the doctrines of criminal law appear to enhance social welfare. This seems true of the basic feature of criminal law that punishment is not imposed on all harmful acts, but instead is usually confined to those that are especially undesirable. (For example, murder is subject to criminal sanctions, but some accidental killing is not.) As we have stressed, when the socially costly sanction of imprisonment is employed, the fault system is desirable because it results in less frequent imposition of punishment than strict liability. Also, the focus on intent in criminal law as a precondition for imposing sanctions may serve to foster deterrence because those who intend to do harm are more likely to conceal their acts, and may be harder to discourage because of the benefits they anticipate. An additional example of a welfare-enhancing doctrine in criminal law concerns attempts. That attempts to do harm are punished is an implicit way of raising the likelihood of sanctions for undesirable acts.

See Also

- ▶ Deterrence (Theory), Economics of
- ▶ Externalities
- ▶ Law, Economic Analysis of
- ▶ Pecuniary Versus Non-pecuniary Penalties

Acknowledgments A. Mitchell Polinsky's research was supported by the John M. Olin Program in Law and Economics at Stanford Law School. Steven Shavell's research was supported by the John M. Olin Center for Law, Economics, and Business at Harvard Law School.

Bibliography

Arlen, J. 1994. The potentially perverse effects of corporate criminal liability. *Journal of Legal Studies* 23: 833–867.

Bebchuk, L., and L. Kaplow. 1992. Optimal sanctions when individuals are imperfectly informed about the probability of apprehension. *Journal of Legal Studies* 21: 365–370.

Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Becker, G., and G. Stigler. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Bentham, J. 1789. An introduction to the principles of morals and legislation. In *The utilitarians*. Garden City: Anchor Books, 1973.

Block, M., and J. Sidak. 1980. The cost of antitrust deterrence: Why not hang a price fixer now and then? *Georgetown Law Journal* 68: 1131–1139.

Bowles, R., and N. Garoupa. 1997. Casual police corruption and the economics of crime. *International Review of Law and Economics* 17: 75–87.

Cooter, R., and D. Rubinfeld. 1989. Economic analysis of legal disputes and their resolution. *Journal of Economic Literature* 27: 1067–1097.

Craswell, R., and J.E. Calfee. 1986. Deterrence and uncertain legal standards. *Journal of Law, Economics, & Organization* 2: 279–303.

Kaplow, L. 1990. Optimal deterrence, uninformed individuals, and acquiring information about whether acts are subject to sanctions. *Journal of Law, Economics, & Organization* 6: 93–128.

Kaplow, L. 1992. The optimal probability and magnitude of fines for acts that definitely are undesirable. *International Review of Law and Economics* 12: 3–11.

Kaplow, L., and S. Shavell. 1994a. Accuracy in the determination of liability. *Journal of Law and Economics* 37: 1–15.

Kaplow, L., and S. Shavell. 1994b. Optimal law enforcement with self–reporting of behavior. *Journal of Political Economy* 102: 583–606.

Kaplow, L., and S. Shavell. 2002. *Fairness versus welfare*. Cambridge, MA: Harvard University Press.

Kornhauser, L. 1982. An economic analysis of the choice between enterprise and personal liability for accidents. *California Law Review* 70: 1345–1392.

Landes, W., and R. Posner. 1975. The private enforcement of law. *Journal of Legal Studies* 4: 1–46.

Landes, W., and R. Posner. 1987. *The economic structure of Tort law*. Cambridge, MA: Harvard University Press.

Levitt, S. 1997. Incentive compatibility constraints as an explanation for the use of prison sentences instead of fines. *International Review of Law and Economics* 17: 179–192.

McAdams, R., and E. Rasmusen. 2007. Norms in law and economics. In *Handbook of law and economics*, vol. 2, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland.

Miceli, T. 1996. Plea bargaining and deterrence: An institutional approach. *European Journal of Law and Economics* 3: 249–264.

Mookherjee, D., and I. Png. 1992. Monitoring vis-à-vis investigation in enforcement of law. *American Economic Review* 82: 556–565.

Mookherjee, D., and I. Png. 1994. Marginal deterrence in enforcement of law. *Journal of Political Economy* 102: 1039–1066.

Newman, H., and D. Wright. 1990. Strict liability in a principal–agent model. *International Review of Law and Economics* 10: 219–231.

Png, I. 1986. Optimal subsidies and damages in the presence of judicial error. *International Review of Law and Economics* 6: 101–105.

Polinsky, A. 1980a. Private versus public enforcement of fines. *Journal of Legal Studies* 9: 105–127.

Polinsky, A. 1980b. Strict liability vs. negligence in a market setting. *American Economic Review* 70: 363–370.

Polinsky, A. 2006a. The optimal use of fines and imprisonment when wealth is unobservable. *Journal of Public Economics* 90: 823–835.

Polinsky, A. 2006b. Optimal fines and auditing when wealth is costly to observe. *International Review of Law and Economics* 26: 232–235.

Polinsky, A., and D. Rubinfeld. 1988. The deterrent effects of settlements and trials. *International Review of Law and Economics* 8: 109–116.

Polinsky, A., and D. Rubinfeld. 1991. A model of optimal fines for repeat offenders. *Journal of Public Economics* 46: 291–306.

Polinsky, A., and S. Shavell. 1979. The optimal tradeoff between the probability and magnitude of fines. *American Economic Review* 69: 880–891.

Polinsky, A., and S. Shavell. 1984. The optimal use of fines and imprisonment. *Journal of Public Economics* 24: 89–99.

Polinsky, A., and S. Shavell. 1992. Enforcement costs and the optimal magnitude and probability of fines. *Journal of Law and Economics* 35: 133–148.

Polinsky, A., and S. Shavell. 1993. Should employees be subject to fines and imprisonment given the existence of corporate liability? *International Review of Law and Economics* 13: 239–257.

Polinsky, A., and S. Shavell. 1998. On offense history and the theory of deterrence. *International Review of Law and Economics* 18: 305–324.

Polinsky, A., and S. Shavell. 1999. On the disutility and discounting of imprisonment and the theory of deterrence. *Journal of Legal Studies* 28: 1–16.

Polinsky, A., and S. Shavell. 2000a. The economic theory of public enforcement of law. *Journal of Economic Literature* 38: 45–76.

Polinsky, A., and S. Shavell. 2000b. The fairness of sanctions: Some implications for optimal enforcement policy. *American Law and Economics Review* 2: 223–237.

Polinsky, A., and S. Shavell. 2001. Corruption and optimal law enforcement. *Journal of Public Economics* 81: 1–24.

Polinsky, A., and S. Shavell. 2007. The theory of public enforcement of law. In *Handbook of law and economics*, vol. 1, ed. A. Polinsky and S. Shavell. Amsterdam: North-Holland.

Posner, R. 1985. An economic theory of the criminal law. *Columbia Law Review* 85: 1193–1231.

Posner, R. 1997. Social norms and the law: An economic approach. *American Economic Review: Papers and Proceedings* 87: 365–369.

Reinganum, J. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78: 713–728.

Rose-Ackerman, S. 1999. *Corruption and government: Causes, consequences and reform*. New York: Cambridge University Press.

Rubinstein, A. 1979. An optimal conviction policy for offenses that may have been committed by accident. In *Applied game theory*, ed. S. Brams, A. Schotter, and G. Schwodiauer. Wurzburg: Physica-Verlag.

Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9: 1–25.

Shavell, S. 1982. On liability and insurance. *Bell Journal of Economics* 13: 120–132.

Shavell, S. 1985. Criminal law and the optimal use of nonmonetary sanctions as a deterrent. *Columbia Law Review* 85: 1232–1262.

Shavell, S. 1987a. The optimal use of nonmonetary sanctions as a deterrent. *American Economic Review* 77: 584–592.

Shavell, S. 1987b. A model of optimal incapacitation. *American Economic Review: Papers and Proceedings* 77: 107–110.

Shavell, S. 1987c. *Economic analysis of accident law.* Cambridge, MA: Harvard University Press.

Shavell, S. 1991. Specific versus general enforcement of law. *Journal of Political Economy* 99: 1088–1108.

Shavell, S. 1992. A note on marginal deterrence. *International Review of Law and Economics* 12: 345–355.

Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36: 255–287.

Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26: 575–612.

Shavell, S. 2002. Law versus morality as regulators of conduct. *American Law and Economics Review* 4: 227–257.

Shleifer, A., and R. Vishny. 1993. Corruption. *Quarterly Journal of Economics* 108: 599–617.

Spier, K. 1997. A note on the divergence between the private and the social motive to settle under a negligence rule. *Journal of Legal Studies* 26: 613–621.

Stigler, G. 1970. The optimum enforcement of laws. *Journal of Political Economy* 78: 526–536.
Sykes, A. 1981. An efficiency analysis of vicarious liability under the law of agency. *Yale Law Journal* 91: 168–206.
Wilde, L. 1992. Criminal choice, nonmonetary sanctions, and marginal deterrence: A normative analysis. *International Review of Law and Economics* 12: 333–344.

Layoffs

John Haltiwanger

Abstract

Layoffs reflect employer-initiated job separations that play an important role in frictional and cyclical unemployment. The relative importance of temporary and permanent layoffs and layoffs themselves has varied over time, and understanding the factors underlying this variation is important for understanding fluctuations in frictional and cyclical unemployment over time. Modern models of labour market dynamics often emphasize the layoffs associated with endogenous job destruction at the firm level induced by the interaction of aggregate and firm-specific shocks.

Keywords

Business cycles; Cyclical unemployment; Hold-up problem; Implicit contracts; Information capital; Labour market search; Layoffs; Search and matching models; Job creation and destruction

JEL Classifications

J63

The term 'layoff' is controversial in itself. For some the term connotes a temporary employer-initiated discharge, for others it represents any employer-initiated discharge that is without prejudice to the worker. The data on layoffs collected by the Bureau of Labor Statistics (BLS) in the United States (see, for example, various issues of the journal *Employment and Earnings*) takes the alternative types of layoffs into account across its firm and household surveys. Layoff data from the BLS survey of firms (the Job Openings and Labor Turnover Survey, JOLTS) provide data on employer- initiated discharges making no distinction as to whether the layoff is temporary or permanent. According to JOLTS, layoffs average about 1.1 per cent of US non-farm employment each month, which is about one-third of all worker separations. The BLS survey of households (the Current Population Survey, CPS) distinguishes between 'temporary layoffs' and 'permanent job losers' in tracking unemployment, where the former are layoffs for which recall is expected within six months and the latter are layoffs where employment ended involuntarily and the workers have begun looking for work. According to the CPS, about 50 per cent of all unemployed are classified as job losers and temporary layoffs account for one-third of the job losers.

The controversy over the terminology is dwarfed by the controversy over the occurrence of layoffs. When General Motors announces that it is laying off 20,000 of its workers indefinitely there is widespread press coverage. This attention is well deserved since substantial variation in layoffs (both temporary and permanent) is frequently observed, and layoffs play an important role in cyclical unemployment. Empirical studies of unemployment (for example, Davis et al. 1996; Bleakley et al. 1999) indicate that the typical increase in unemployment during a business cycle slump is primarily due to an increase in employer-initiated discharges, that is, layoffs. For example, in the sharp 1982 recession in the USA, the fraction of the unemployed due to job loss peaked at 63 per cent while in the 2001 recession this fraction peaked at 56 per cent.

The increase in layoff unemployment during recessions is closely tied to the increase in gross job destruction in recessions. Davis et al. (1996) show that job destruction rises substantially during recessions and is increasingly driven by establishments contracting substantially (for example, with contractions greater than 25 per cent). In turn, Davis et al. (2006) show that establishments

that are contracting intensively use layoffs as the primary means of contraction.

The structure of temporary and permanent layoffs over the cycle has varied over time. Groshen and Potter (2003) show that, in the four recessions in the USA between 1967 and 1990, both temporary layoff and permanent layoff unemployed surged in each of the recessions. However, starting with the 1990–1 recessions, temporary layoffs have played a much smaller role and the rise in job loss has been driven almost entirely by permanent layoffs.

The theory of layoff unemployment has evolved with the relative importance of temporary versus permanent layoffs. Given the important role for temporary layoffs in the 1970s, the so-called 'implicit contract models' (see, for example, Azariadis 1975; Baily 1974; Burdett and Mortensen 1980) were developed during that time to help account for the role of temporary layoffs. The temporary layoff models provide a basis for understanding how in a long-term employer-employee relationship it may be optimal for firms and workers to use temporary layoffs to respond to transitory shocks. However, the increased understanding and role of permanent job destruction and associated permanent job loss has pushed theoretical developments in new directions.

Recent theories that incorporate the evidence on permanent job destruction adopt the premise that the economy is subject to a continuous stream of allocative shocks – shocks that cause idiosyncratic variation in profitability among job sites and worker–job matches (see Davis and Haltiwanger 1999; Mortensen and Pissarides 1999; Shimer et al. 2005 for an extensive survey of these theories). The continuous stream of allocative shocks generates the large-scale job and worker reallocation observed in the data. To explicitly model the job and worker reallocation process, these theories incorporate heterogeneity among workers and firms along one or more dimensions. Various theories also emphasize search costs, moving costs, sunk investments and other frictions that impede or otherwise distort the reallocation of factor inputs. The combination of frictions and heterogeneity gives rise to potentially important roles for allocative shocks and the reallocation process in aggregate economic fluctuations.

Theories of cyclical fluctuations in job and worker flows with such reallocation frictions can be classified into two broad types. One type treats fluctuations over time in the intensity of allocative shocks as an important driving force behind aggregate fluctuations and the pace of reallocation activity. A second type maintains that while allocative shocks and reallocation frictions are important, aggregate shocks drive business cycles and fluctuations in the pace of worker and job reallocation. Although different in emphasis, the two types of theories offer complementary views of labour market dynamics and business cycles, and both point toward a rich set of interactions between aggregate fluctuations and the reallocation process.

One can think of allocative shocks as events that alter the closeness of the match between the desired and actual characteristics of labour and capital inputs. Adverse aggregate consequences can result from such events because of the time and other costs of reallocation activity. In considering this view, it is important to emphasize that allocative shocks affect tangible inputs to the production process (labour and physical capital) and intangible inputs. These intangible inputs include the information capital embodied in an efficient sorting and matching of heterogeneous workers and jobs, knowledge about how to work productively with co-workers, knowledge about suitable locations for particular business activities and about idiosyncratic attributes of those locations, the information capital embodied in long-term customer–supplier and debtor–creditor relationships, and the organization capital embodied in sales, product distribution and job-finding networks. These remarks make clear why the economic adjustments to these shocks are often costly and time consuming. It follows that sharp time variation in the intensity of allocative shocks can cause large fluctuations in gross job flows and in turn unemployment dynamics and layoffs in particular.

The connection between cyclical fluctuations in job destruction and layoffs may also stem from

responses to adverse aggregate shocks. An adverse aggregate shock can push many declining and dying plants over an adjustment threshold. During boom times, a firm may choose to continue operating a plant that fails to recover its long-run average cost, because short-run revenues exceed short-run costs, or because of a sufficiently large option value to retaining the plant and its work force. A closely related mechanism emphasizes the changes in the incentives for reallocation over the cycle. The reallocation of specialized labour and capital inputs involves forgone production due to lost work time (for example, unemployment or additional schooling), worker retraining, the retooling of plant and equipment, the adoption of new technology, and the organization of new patterns of production and distribution. On average across firms and workers, the value of forgone production tends to fluctuate procyclically, rising during expansions and falling during recessions. This cyclical pattern generates incentives for both workers and firms to concentrate costly reallocation activity during recessions, when the opportunity cost of the resulting forgone production is relatively low. This mechanism is highlighted in the models of Davis and Haltiwanger (1999), Mortensen and Pissarides (1994), and Caballero and Hammour (1994).

A key question is whether the cyclical fluctuations in job destruction and layoffs reflect efficient or inefficient responses to shocks. Caballero and Hammour (1996) highlight the potential for labour markets to malfunction because of appropriability or hold-up problems. These problems arise whenever investment in a new production unit or the formation of a new employment relationship involves some degree of specificity for workers or employers, and there are difficulties in writing or enforcing complete contracts. In their model, Caballero and Hammour (1996) show that efficient restructuring involves synchronized job creation and destruction and relatively little unemployment. In contrast, the inefficient equilibrium restructuring process that emerges under incomplete contracts involves the decoupling of creation and destruction dynamics and relatively large unemployment responses to negative shocks. As discussed in Mortensen and Pissarides (1999), appropriability problems arise naturally in many search and matching models.

Malcomson (1999) provides a broad discussion of hold-up problems in the labour market.

Overall, understanding layoffs requires understanding of the underlying dynamics of job and worker reallocation. New theories and new data sets have emerged that provide a rich new perspective on the dynamics of the labour market at the micro level and in turn the implications of these dynamics for aggregate fluctuations. Much work remains to be done on both theoretical and empirical questions, particularly on understanding the role of market imperfections in these dynamics. Along these lines, one continuing open question is not only to understand the driving forces of job loss but also the closely related forces of the job gains.

After all, the loss of a job has much lower costs to the individual and the economy if the worker in question moves quickly to another job.

L

See Also

- ▶ Natural Rate of Unemployment
- ▶ Search Models of Unemployment

Bibliography

Ashenfelter, O., and D. Card. 1999. *Handbook of labor economics*. Amsterdam: North-Holland.

Azariadis, C. 1975. Implicit contracts and underemployment equilibria. *Journal of Political Economy* 83: 1183–1202.

Baily, M.N. 1974. Wages and employment under uncertain demand. *Review of Economic Studies* 41: 37–50.

Bleakley, H., A. Ferris, and J. Fuhrer. 1999. New data on worker flows during business cycles. *Federal Reserve Bank of Boston Review* July/August, 49–76.

Burdett, L., and D. Mortensen. 1980. Search, layoffs, and labor market equilibrium. *Journal of Political Economy* 88: 652–672.

Bureau of Labor Statistics. Current population survey. Available at http://www.census.gov/cps. Accessed 8 Mar 2007.

Bureau of Labour Statistics. Job openings and labor turnover survey. Available at http://www.bls.gov/jlt. Accessed 8 Mar 2007.

Caballero, R., and M. Hammour. 1994. The cleansing effect of recessions. *American Economic Review* 84: 1350–1368.

Caballero, R., and M. Hammour. 1996. On the timing and efficiency of creative destruction. *Quarterly Journal of Economics* 111: 805–852.
Davis, S.J., and J. Haltiwanger. 1999. Gross job flows. In Ashenfelter and Card (1999).
Davis, S.J., J. Haltiwanger, and S. Schuh. 1996. *Job creation and destruction*. Cambridge, MA: MIT Press.
Davis, S.J., J. Faberman, and J. Haltiwanger. 2006. The flow approach to labor markets: New data sources and micro-macro links. *Journal of Economic Perspectives* 20 (3): 3–26.
Groshen, E., and S. Potter. 2003. Has structural change contributed to a jobless recovery? *Current Issues in Economics and Finance* 9 (8): 1–7.
Malcomson, J.M. 1999. Individual employment contracts. In Ashenfelter and Card (1999).
Mortensen, D.T., and C. Pissarides. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61: 397–415.
Mortensen, D.T., and C. Pissarides. 1999. New developments in models of search in the labor market. In Ashenfelter and Card (1999).
Shimer, R., R. Rogerson, and R. Wright. 2005. Search theoretic models of the labor market. *Journal of Economic Literature* 43: 959–988.
U.S. Department of Labor. 2007. *Employment and earnings, various issues*. Washington, DC: Government Printing Office.

Layton, Walter Thomas (1884–1966)

Murray Milgate and Alastair Levy

Variously occupied as academic economist, civil servant, economic journalist, economic adviser and newspaper magnate, Layton was a product of Cambridge before World War I. Born at Chelsea on 15 March 1884, he was educated at Westminster City School, University College, London (1901–4) and Trinity College, Cambridge. He came up to Cambridge in 1904 to read economics (the new Tripos having been established only the year before) and took a double First. In 1908 Pigou engaged him as an assistant lecturer (paying him out of his own pocket, much as Marshall had done with Pigou), in 1909 he was elected into a fellowship at Caius College, and in 1911 he was appointed University Lecturer. World War I took him out of Cambridge, first to the Board of Trade and then to the Ministry of Munitions. In 1921 he succeeded Hartley Withers as editor of the *Economist* where he remained until 1938, though his association with it continued and when he died in 1966 he was still vice-chairman of the board.

Layton's first academic publication appeared in the *Economic Journal* for March 1905 – a not inconsiderable feat for a first year undergraduate at Cambridge. His first article as an 'economic journalist' appeared in the *Economist* in 1908 while he was still a Cambridge don. In this period, however, his two most significant publications were *An Introduction to the Study of Prices* (1912) and *The Relations of Capital and Labour* (1914). The former was an analysis of fluctuations in the general level of prices in the 19th century (and their impact on living standards and the distribution of income), and movements of money and real wages. Layton's interest in applied economics had no doubt been stimulated by his contact as a student (both at University College and Cambridge) with C.P. Sanger, and it was undoubtedly cemented by the sound advice given to him by Marshall in 1910 to the effect that if he wished to do two things at once (that is, to be both an academic economist and an economic journalist), he should ensure that his research for the one had external effects for the other (letter from Marshall, 1910, quoted in Hubback 1985, p. 32). Applied research into prices and movements in money and real wages was just the thing His first lectures at Cambridge were on the problems of industry and labour, for three years from 1909 he gave the Newmarch lectures at University College on statistics, at the *Economist* he 'set to work to revise the fifty year old Price Index number' (quoted in Hubback 1985, p. 30), and he gave classes at the Workers Educational Association on applied economics. When the war came, he moved to the Board of Trade (at the request of Beveridge) where he supervised the census of employment, and then transferred to the Ministry of Munitions under Lloyd George where he was appointed Director of Requirements and Statistics.

In 1916, Layton joined Lord Milner's mission to Petrograd – his colleagues returned from Russia

fully convinced there would be no revolution till after the war, but Sir Walter Layton was perhaps an exception. When asked . . . 'Are they keen on war?', he replied, 'No, they are much too busy thinking of the coming revolution' (Lloyd George 1935, p. 942). Layton's wartime services were rewarded in 1919 when he was made a Companion of Honour. His success at the Ministry of Munitions led to a part in the establishment, and later a directorship, of the financial section of the League of Nations. After the war, Layton made the relatively easy transition to full-time economic journalism. His career from this time on becomes of less interest to economists (though probably of greater interest for students of the history of Fleet Street), save for two episodes.

Layton was always attracted to Liberal politics, and after World War I this attachment took on a more concrete form. As chairman of the Executive Committee of the Liberal Party's inquiry into the post-war British economy, he was primarily responsible for the organization and publication of the famous Liberal *Yellow Book* of 1928. He was chairman of the Statistics Subcommittee and worked under Keynes for the Finance and Industry Sub-committee. The second episode concerns his role in the Ministries of Supply and Production, to the first of which he was appointed by Churchill in 1940. He rose to become chairman (1942–3) of the Joint War Production Staff which he helped to create, and conducted negotiations in Washington on behalf of the British government to secure material assistance from the Americans for the conduct of the war in its early years. Once again, he worked in concert with Keynes as he had done in the first war.

Layton was created Baron Layton of Dane Hill in 1947, and was deputy leader of the Liberals in the House of Lords from 1952 to 1955. He died in the late winter of 1966 after contracting pneumonia.

Selected Works

1912. *An introduction to the study of prices, with special reference to the history of the nineteenth century.* London: Macmillan.

1914. *The relations of capital and labour.* London/Glasgow: Collins.

Bibliography

Hubback, D. 1985. *No ordinary press baron: A life of Walter Layton*. London: Weidenfeld & Nicolson.

Lloyd George, D. 1935. *War memoirs*. London: Oldhams Press.

Le Chatelier Principle

Eugene Silberberg

Abstract

In the field of economics, the Le Chatelier principle refers to the differences in the responses of decision variables to changes in parameters when additional constraints are imposed on the system. In the context of demand theory, for example, the Le Chatelier principle is the 'second law of demand', that demand curves are more elastic in the long run than in the short run. In many models, additional constraints reduce the absolute response of a decision variable to a change in a parameter.

Keywords

Comparative statics; Conjugate pairs theorem; Envelope theorem; Le Chatelier principle; Parameter values; Samuelson P.

JEL Classifications

D11

Henri Louis Le Chatelier was a French chemist born in Paris in 1850. In 1884, he offered the following observation:

> Any system in stable chemical equilibrium, subjected to the influence of an external cause which tends to change either its temperature or its condensation (pressure, concentration, number of

> molecules in unit volume), either as a whole or in some of its parts, can only undergo such internal modifications as would, if produced alone, bring about a change of temperature or of condensation of opposite sign to that resulting from the external cause. (Oliver and Kurtz 1992)

Later writers produced a more heuristic simplification: 'If the external conditions . . . are altered, the equilibrium . . . will tend to move in such a direction so as to oppose the change in external conditions' (Fermi 1937, p. 111, cited in Samuelson 1949, p. 639), or even more simply: if a stress is applied to a system at equilibrium, then the system readjusts, if possible, to reduce the stress. The Le Chatelier principle is a firmly established proposition in classical thermodynamics, though its verbal statement is somewhat vague in operational content. In the field of economics, the law of demand, which states that as a price increases, ceteris paribus, consumers will decrease their consumption of that good, is in fact a direct application of the Le Chatelier principle. Consumers (or firms) mitigate the adverse effects of the price increase by utilizing less of that good or input.

Following up a suggestion by his professor and mentor E.B. Wilson at Harvard, Paul Samuelson showed that this principle was a simple application of maximizing behaviour (see especially Samuelson 1949, 1960a, 1974.) Moreover, physicists and economists – among economists, principally Samuelson – came to realize that the Le Chatelier principle was being used to describe two separate phenomena. The first referred to first-order changes in response to a change in a parameter value, such as a price. The second, which is what the Le Chatelier principle is now generally understood to mean, refers to *differences* in the changes as additional constraints are imposed on the system.

The General Case

First-Order Effects

The most general comparative statics model with explicit maximizing behaviour is *maximize* $y = f(x, \alpha)$ subject to $g(x, \alpha) = 0$, where x $(x_1, \ldots, x_n)$ is a vector of decision variables, $\alpha = (\alpha_1, \ldots, \alpha_m)$ is a vector of parameters (though for simplicity, we treat α 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). Thus we restrict the analysis to models of the form

$$maximize \;\; y = f(x, \alpha) \tag{1}$$

$$subject\ to\ g(x) = 0 \tag{2}$$

Since it has no effect on the analysis to follow, we consider the case of only one external constraint. Also, parameters β, which enter the constraint but which do not enter the objective function, also do not affect the analysis, and hence we suppress them in the notation. The Lagrangian for this model is $L = f(x, \alpha) + \lambda g(x)$ producing the necessary first-order conditions (NFOC)

$$L_i = f_i(x, \alpha) + \lambda g_i(x) = 0 \quad i = 1, \ldots, n \tag{3}$$

$$L_\lambda = g(x) = 0 \tag{4}$$

Assuming the sufficient second-order conditions hold, we can in principle 'solve' for the $n+1$ explicit choice functions $x = x^*(\alpha)$ and $\lambda^*(\alpha)$. Of course, since these choice functions are the result of solving the NFOC simultaneously, each individual *xi* is a function of *all* the parameters, not just the ones which appear in *Li*.

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, subject to the constraint. 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

$$maximize \;\; F(x, \alpha) = f(x, \alpha) - \phi(\alpha) \tag{5}$$

$$subject\ to\ g(x) = 0 \tag{6}$$

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 Lagrangian for this model is

$$L = f(x, \alpha) - \phi(\alpha) + \lambda g(x) \tag{7}$$

The first-order conditions with respect to x are the same as in the original model. With respect to α, the NFOC yield the famous 'envelope theorem'

$$L_\alpha = f_\alpha - \phi_\alpha = 0 \tag{8}$$

When α enters the constraint also, we get the envelope theorem in its most general form,

$$\phi_\alpha = L_\alpha = f_\alpha + \lambda g_\alpha \tag{8a}$$

Importantly, however, since we have restricted the model so that the parameters α do not enter the constraint, the primal-dual model is *an unconstrained maximization in* α. Hence in the α dimensions, the second-order conditions are simply

$$F_{\alpha\alpha} = f_{\alpha\alpha} - \phi_{\alpha\alpha} \leq 0 \tag{9}$$

This inequality says that in the α dimensions, f is relatively more concave than ϕ. This is the fundamental geometrical property that underlies all comparative statics relationships and also the 'second-order' Le Chatelier relationships.

The NFOC (8) are identities when $x = x^*$. That is,

$$\phi_\alpha(\alpha) \equiv f_\alpha(x^*(\alpha), \alpha) \tag{10}$$

Differentiating with respect to α,

$$\phi_{\alpha\alpha} \equiv \sum_1^n f_{\alpha i} \frac{\partial x_i^*}{\partial \alpha} + f_{\alpha\alpha} \tag{11}$$

Rearranging terms, using (9) and invariance to the order of differentiation,

$$\phi_{\alpha\alpha} - f_{\alpha\alpha} \equiv \sum_1^n f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0 \tag{12}$$

This is the fundamental relation of comparative statics. From it, we can derive Samuelson's famous 'conjugate pairs' theorem, namely, that refutable implications occur in maximization models when and only when a parameter enters one and only one first-order condition. For in that case, where say α enters only $Li = 0$, $f_{j\alpha} \equiv 0$, $j \neq i$, and so (12) reduces to one term:

$$f_{i\alpha} \frac{\partial x_i^*}{\partial \alpha} \geq 0 \tag{13}$$

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). These relationships constitute the 'first-order' Le Chatelier effects. Note that these results are identical to those in models with no constraints at all, or with multiple constraints, as long as those constraints do not contain the parameter that is changing.

L

Second-Order Effects

Suppose now the NFOC hold at the parameter value α^0 and consider now the imposition of an additional constraint, $h(x) = 0$, with the important restriction that this constraint does not change the original equilibrium, for example, a constraint holding some input fixed at the previous profit maximizing level. Then the new NFOC are solved for new explicit choice functions, $x_i = x_i^s(\alpha)$, where the superscript 's' stands for 'short run'. Substituting these short run choice functions into the objective function produces a new indirect objective function, $\psi(\alpha)$. Since the new constraint did not disturb the equilibrium, $\psi(\alpha^0) = \phi(\alpha^0)$ at that point. However, since the objective function is now more constrained, for $\alpha \neq \alpha^0$, $\psi(\alpha) \leq \phi(\alpha)$. Thus the function $G(\alpha) = \psi(\alpha) - \phi(\alpha)$ has an unconstrained maximum (of zero) at $\alpha = \alpha^0$. The NFOC are

$$G_\alpha(\alpha) = \psi_\alpha(\alpha) - \phi_\alpha(\alpha) = 0 \tag{14}$$

We note that $\psi_\alpha(\alpha) = \phi_\alpha(\alpha) = f_\alpha$ using the same analysis leading to Eq. 8, since α appears

in neither constraint. The second-order conditions are

$$G_{\alpha\alpha}(\alpha) = \psi_{\alpha\alpha}(\alpha) - \phi_{\alpha\alpha}(\alpha) \leq 0 \qquad (15)$$

That is, the more constrained indirect objective function $\psi(\alpha)$ is tangent to $\phi(\alpha)$ at $\alpha = \alpha^0$, but it is relatively more concave, or less convex. Using $\psi_\alpha(\alpha) \equiv \phi_\alpha(\alpha) \equiv f_\alpha$ expressed as identities, and proceeding as in Eqs. (10) through (12), inequality (15) yields the general second-order Le Chatelier effects:

$$\sum_1^n f_{i\alpha}\left(\frac{\partial x_i^*}{\partial x} - \frac{\partial x_i^s}{\partial x}\right) \geq 0 \qquad (16)$$

In the empirically important case where α enters only the ith first-order condition, this summation reduces to one term, producing

$$f_{i\alpha}\frac{\partial x_i^*}{\partial \alpha} \geq f_{i\alpha}\frac{\partial x_i^s}{\partial \alpha} \qquad (17)$$

Thus $\partial x_i^*/\partial \alpha \geq \partial x_i^s/\partial \alpha \geq 0$ when $f_{i\alpha} > 0$, and $\partial x_i^*/\partial \alpha \leq \partial x_i^s/\partial \alpha \leq 0$ when $f_{i\alpha} < 0$. In either case, $\left|\partial x_i^*/\partial \alpha\right| \geq \left|\partial x_i^s/\partial \alpha\right|$.

Examples

Profit Maximization

Consider the profit-maximization model *maximize* $\pi = f(x, w, p) = p\theta(x_1 \ldots, x_n) - \sum w_i x_i$. Each parameter w_i enters only the ith NFOC, and $f_{x_i w_i} = -1$, so that (13) yields the negative slope property $\partial x_i/\partial w_i \leq 0$.

Moreover, (17) yields, in addition, for any additional constraint (not involving w_i) imposed on the initial equilibrium,

$$\frac{\partial x_i^*}{\partial w_i} \leq \frac{\partial x_i^s}{\partial w_i} \leq 0 \qquad (18)$$

The 'long-run' factor demand functions are more elastic than any short-run factor demands defined as above.

In the case where the additional constraint is simply $x_n = x_n^0$, an analysis based on 'conditional demands' (Pollak 1969) is available. If we substitute this constraint directly into the objective function, the 'short-run' demand functions are $x_i = x_i^s(w_1, \ldots w_{n-1}, p, x_n^0)$. These functions are related to the long-run demands by the identity

$$x_i^*(w_1, \ldots, w_n, p) \times \equiv x_i^s\left(w_1, \ldots w_{n-1}, p, x_n^*(w_1, \ldots, w_n, p)\right) \qquad (19)$$

Differentiating both sides of this identity with respect to w_i and w_n,

$$\frac{\partial x_i^*}{\partial w_i} \equiv \frac{\partial x_i^s}{\partial w_i} + \frac{\partial x_i^s}{\partial x_n^0}\frac{\partial x_n^*}{\partial w_i} \qquad (20)$$

$$\frac{\partial x_i^*}{\partial w_n} \equiv \frac{\partial x_i^s}{\partial x_n^0}\frac{\partial x_n^*}{\partial w_n} \qquad (21)$$

Substituting (21) into (20) and using a well-known reciprocity condition yields

$$\frac{\partial x_i^*}{\partial w_i} \equiv \frac{\partial x_i^s}{\partial w_i} + \frac{\left(\partial x_i^*/\partial w_n\right)^2}{\partial x_n^*/\partial w_n} \qquad (22)$$

Since the last term in (22) is negative, we get the Le Chatelier result (18).

Cost (Expenditure) Minimization

The cost functions in production theory are derived from the model, *minimize* $C = \sum w_i x_i$ subject to $f(x_1, \ldots, x_n) = y$, where y is now a parameter, that is, it is an arbitrary fixed level of output. This model is directly related to the profit maximization model. Write the profit maximization model as *maximize* $py - \sum w_i x_i$ *subject to* $f(x_1, \ldots, x_n) = y$. When output y is a variable, this model is the profit- maximization model. If y is parametric, it is the constrained cost minimization model. Thus we see that the cost minimization model is the profit maximization model with an added constraint. Denoting the factor demands derived from cost minimization as $x_i = x_i^y(w_1, w_n, y)$, we apply (13) and (17) to derive $\partial x_i^*/\partial w_i \leq \partial x_i^y/\partial w_i \leq 0$. The profit maximizing factor demand function, which incorporates an output effect, is always more elastic with respect to its own price than the constant output factor demand functions, regardless of whether the output

effect is positive or negative. We can also show by this method that, if another constraint is imposed on the factors, these cost-minimizing demand functions become less elastic. When the additional constraint takes on the form of holding some factor fixed, as in the above profit-maximization model, a similar conditional demand process is available (see Silberberg and Suen 2000).

Marginal Cost Functions

Many – perhaps most – important economic models incorporate a constraint of the form $g(x_1, \ldots, x_n) = k$. The cost minimization model is an example; so are the various two-factor two-good models in which endowment levels are fixed. The Lagrangian for the cost minimization model is $L = \sum w_i x_i + \lambda(y - f(x_1, \ldots, x_n))$. The indirect objective function is the cost function $C = C^*(w_1, \ldots, w_n, y)$. The envelope theorem (8a) identifies $\lambda^*(w_1, \ldots, w_n, y)$ as the marginal cost function: $C^*_y = \lambda^*$. We know from the above comparative statics discussion that cost minimization does not imply a sign for the slope of the marginal cost function, that is, $\partial\lambda^*/\partial y/0 \rightarrow \partial\lambda^*/\partial\lambda^*/\partial y \gtrless 0$. Nonetheless, we can still derive a Le Chatelier result for the marginal cost function.

Adding a new constraint $h(x) = 0$ to the cost minimization model consistent with the original equilibrium produces a new 'short run' cost function $C^s(w_1, \ldots w_n, y)$. Since this is more constrained than C^*, it must be the case that $C^* \leq C^s$, but the two are equal at the original equilibrium. Thus the function $F = C^* - C^s$ has an unconstrained maximum (of zero) with respect to all the parameters, and in particular, y. Thus $F_y = C^*_y - C^s_y = 0 \text{ and } F_{yy} = C^*_{yy} - C^s_{yy} \leq 0$. But this latter inequality is $\partial\lambda^*/\partial y \leq \partial\lambda^s/\partial y$. That is, the long-run marginal cost function either falls faster or rises slower than the short-run marginal cost function. This is the mathematical foundation for the famous article by Viner (1932) and his draftsman Wong that started it all.

Extensions

The Le Chatelier principle is a local result. Even with the usual sufficient second-order conditions, if some price changes by a finite amount, it is not an implication of the model that the long-run effects are absolutely larger than the short-run effects.

However, Milgrom and Roberts (1996) showed, using lattice theory, that, for example, for the profit-maximizing firm model, if all the cross-partials of the production function are everywhere non-negative, the Le Chatelier results hold in the large. A few years later, Suen, Silberberg and Tseng (2000) provided an easier proof of this result, showing also that the global Le Chatelier result held when the factors of production and the fixed factor do not switch from being substitutes to being complements (or vice versa) over the relevant price range.

Samuelson (1960a) analysed Le Chatelier phenomena for equilibrium systems not resulting from an explicit maximization hypothesis, using the 'well-known' theorem of reciprocal determinants of Jacobi. (I used to joke to my classes that the theorem was well-known to Jacobi and to Samuelson.) Lady and Quirk (2004) have analysed non-maximizing systems using a theory of cycles in determinants; they prove the Le Chatelier principle applies to systems identified by Morishima (1952), which allows substitutes and complements.

See Also

- Comparative Statics
- Envelope Theorem

Bibliography

Fermi, E. 1937. *Thermodynamics*, 111. New York: Dover Publications.

Lady, G., and J. Quirk. 2004. The scope of the Le Chatelier principle. Online. Available at http://optima-com.com/LeChat/The%20Scope%20of%20the%20LeChatelier%20Principle.doc. Accessed 1 Feb 2006.

Milgrom, P., and J. Roberts. 1996. The Le Chatelier principle. *American Economic Review* 86: 173–179.

Morishima, M. 1952. On the laws of change in the price system in an economy which contains complementary commodities. *Osaka Economics Papers* 1: 101–113.

Oliver, J. and J. Kurtz. 1992. Henri Louis Le Chatelier, a man of principle. Woodrow Wilson Fellowship Foundation.

Pollak, R. 1969. Conditional demand functions and consumption theory. *Quarterly Journal of Economics* 83: 60–78.
Samuelson, P. 1949. *The Le Chatelier principle in linear programming. RAND Corporation Monograph*. Reprinted in *The collected scientific papers of Paul Samuelson*, vol. 1, ed. J. Stiglitz. Cambridge, MA: MIT Press, 1966.
Samuelson, P. 1960a. Structure of a minimum equilibrium system. In *Essays in economics and econometrics: A volume in honor of Harold Hotelling*, ed. R. Pfouts. Chapel Hill: University of North Carolina Press.
Samuelson, P. 1960b. An extension of the le Chatelier principle. *Econometrica* 28: 368–379. Reprinted in *The collected scientific papers of Paul Samuelson*, vol. 1, ed. J. Stiglitz. Cambridge, MA: MIT Press, 1966.
Samuelson, P. 1974. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Silberberg, E., and W. Suen. 2000. *The structure of economics*. 3rd ed. New York: Irwin/McGraw-Hill.
Suen, W., E. Silberberg, and P. Tseng. 2000. The Le Chatelier principle: The long and the short of it. *Economic Theory* 16: 471–476.
Viner, J. 1932. Cost curves and supply curves. *Zeitschrift fur Nationalokonomie* 3: 23–46. Reprinted in American Economic Association, *Readings in price theory*. Chicago: Richard D. Irwin, 1952.

Le Trosne, Guillaume François (1728–1780)

Peter Groenewegen

Keywords
Le Trosne, G. F.; Physiocracy; Single tax; Value

JEL Classifications
B31

French lawyer and economist. Born in Orléans, Le Trosne studied natural law philosophy with Pothier in preparation for work as a magistrate. In 1753 he was appointed Royal Councillor at the Orléans Presidial Court, whence he retired in 1773. Le Trosne joined the Physiocrats in 1764 by publishing a book defending the free trade in grain (1765) and articles in *Ephémérides* and other journals. His major economic work, *De l'ordre social*, appeared in 1777, its second volume, *De l'intérêt social*, having major economic content with its discussion of value, circulation, money, industry, and domestic, foreign and colonial trade, partly by way of criticism of Condillac's (1776) anti-physiocratic views on these subjects. Le Trosne died in Paris in 1780.

De l'ordre social sets out the laws required for good government designed to ensure and enhance the reproduction of subsistence and wealth. Two major laws are identified. The first demands freedom for economic activity and security of property (Le Trosne 1777a, p. 38). The second seeks to secure sufficient government revenue to defray public expenses in providing not only security of property and defence but also public works in communication and transport most favourable to reproduction (1777a, p. 122). The second law entails an appropriate tax system ensured by gradual implementation of the single tax on net product (1777a, p. 147). The remaining discourses of the first volume develop the absolute necessity of these laws from historical examples and from their undesirable consequences when transgressed. Constitutional issues of good government defended in part by standard physiocratic arguments in lengthy footnotes (for example, on luxury 1777a, pp. 214–19, and free trade, pp. 347–50) form the thrust of the argument in the first volume.

Le Trosne's second volume (1777b) is particularly noted for its theory of value (Meek 1962, p. 389, n. 1), which distinguishes its various determinants such as usefulness, tastes, relative scarcity and competition but which identifies necessary expenses of production as the major influence on value, hence the name fundamental price (pp. 503–4). To analyse value effects on production and wealth Le Trosne distinguishes various value forms linking, for example, the excess of the price received for produce by the farmer over costs, to accumulation and the increase of wealth. Other roles for these complex value relationships are illustrated in Le Trosne's perceptive discussions of exchange, money, circulation, the sterility of industry and the benefits

of trade for an agricultural nation. This analysis clearly confirms the value foundations of physiocratic theory, crystallized in his demonstration of the special productivity of agriculture by means of a simple example where all payments are assumed to be in kind ('*en nature*'), thereby demonstrating the inaccuracy of interpretations which neglect the sophisticated physiocratic value analysis (p. 590).

Selected Works

1765. *La liberté du commerce des grains, toujours utile et jamais nuisible*. Paris.
1777a. *De l'ordre social*. Paris. Reprinted. Munich: Kraus, 1980.
1777b. *De l'intérêt social, par rapport à la valeur, à la circulation, à l'industrie, & au commerce intérieur & extérieur*. Paris. Reprinted. Munich: Kraus, 1980.

Bibliography

de Condillac, E.B. 1776. *Le commerce et le gouvernment considérés relativement l'un à l'autre*. Paris.
Meek, R.L. 1962. *The economics of physiocracy*. London: George Allen & Unwin.

Leads and Lags

Olivier Jean Blanchard

The notion that an economic variable leads or lags another variable is an intuitive and simple notion. Nevertheless, it has proven difficult to go from this intuitive notion to a precise, empirically testable, definition.

The first attempt was made by Burns and Mitchell in their work on business cycles (1946). Their interest was in characterizing whether individual variables led or lagged the cycle. Their approach was roughly as follows. It was first to divide time into separate business cycles, then to look at the deviation of each variable from its mean value during each cycle and finally to average across cycles. If the variable reached its maximum – or minimum – value on average before the peak of the cycle, the variable was said to lead the cycle; if it reached its maximum or minimum after the peak, it was said to lag the cycle. Following the same line, Burns and Mitchell constructed an index of leading indicators, composed of a dozen series. The series were chosen by taking into account several criteria, the most relevant – for our purposes – being timing at troughs and peaks. This index is still in use today and is regularly published by the US Department of Commerce.

The implicit definition offered by Burns and Mitchell of a leading variable is quite sophisticated, being a relation between timings of turning points between series. It is also partly judgemental, as the procedure used by Burns and Mitchell implies finding business cycles in the data, deciding on what average behaviour of a series in a typical cycle is and so on. The development of time series methods has led to a quest for a less judgemental and more easily testable definition; this has led to tighter, testable but less sophisticated definitions. The focus has shifted from looking at the relation between two time series at specific points, such as turning points in the Burns–Mitchell work, to looking at characteristics of the joint behaviour of the two time series in general, throughout the business cycle for example. A simple definition is the following: a variable may be said to lead another series if it tends to move – increase, decrease,. . . – before this other series. This still vague statement can be given precise statistical meaning. For example, a series can be said to lead another in the business cycle if the phase difference cross spectral density of the two series is positive at business cycle frequencies. This definition does not capture exactly the same thing as the Burns–Mitchell definition which focuses on particular points, namely turning points, rather than on specific frequencies. But it is close and is easily testable.

This definition, as well as the Burns–Mitchell definition, partly fails however to capture what the intuitive notion of a leading variable is about. In this intuitive notion is the idea that a variable

which leads another contains information about future values of the other that one could not obtain by just looking at current and past values of this other variable. For example, the formal definition given above implies that, if one looked at two sinusoids with close peaks, one would define the one which peaks first as leading the other; it is clear however that the leading sinusoid would contain no information about the other. Thus, one is led to look for a definition which takes into account this notion of additional information.

Such a definition is the following. A variable x leads a variable y if in the following regression,

$$y = a_1 y(-1) + \ldots + a_n y(-n) + b_1 x(-1) + \ldots + b_n x(-n) + e$$

the set of coefficients $(b_1, \ldots, b_n)$ is significant.

This definition captures the notion that x helps predict y, even when one looks at the history of y. If the set of b's is significant, x is also said to cause or 'Granger-cause' y. This definition is easily testable and has become widely accepted. Interestingly, most of the components of the index of leading indicators turn out to lead industrial production in that sense. They help forecast industrial production.

It should be clear, and this is partly obscured by the use of the word causality in this context, that a variable may lead another one, not because it affects it with a lag, but just because it reflects information about its future values. The stock market for example is a leading indicator; this may be because stock prices incorporate information about the future of the economy which one cannot obtain by looking only at current and past values of output, not because stock prices affect economic activity. It may also be a combination of both effects.

See Also

- ▶ Adjustment Costs
- ▶ Indicators
- ▶ Multivariate Time Series Models

Bibliography

Burns, A.F., and W.G. Mitchell. 1946. *Measuring business cycles*. New York: Columbia University Press.

Learning and Evolution in Games: Adaptive Heuristics

H. Peyton Young

Abstract

A 'heuristic' is a method or rule for solving problems; in game theory it refers to a method for learning how to play. Such a rule is 'adaptive' if it is directed towards higher payoffs and is reasonably simple to implement. This article discusses a variety of such rules and the forms of equilibrium that they implement. It turns out that even sophisticated solution concepts, like subgame perfect equilibrium, can be achieved by relatively simple and intuitive methods.

Keywords

Adaptive heuristics; Commitment; Correlated equilibrium; Learning; Nash equilibrium; Probability; Regret; Repeated games; Strategic learning; Subgame perfection

JEL Classifications

C7

'Adaptive heuristics' are simple behavioural rules that are directed towards payoff improvement but may be less than fully rational. The number and variety of such rules are virtually unlimited; here we survey several prominent examples drawn from psychology, computer science, statistics and game theory. Of particular interest are the informational inputs required by different learning rules and the forms of equilibrium to which they lead. We shall begin by considering very primitive heuristics, such as reinforcement

learning, and work our way up to more complex forms, such as hypothesis testing, which still, however, fall well short of perfectly rational learning.

One of the simplest examples of a learning heuristic is *cumulative payoff matching*, in which the subject plays actions next period with probabilities proportional to their cumulative payoffs to date. Specifically, consider a finite stage game G that is played infinitely often, where all payoffs are assumed to be strictly positive. Let a_{ij} (t) denote the cumulative payoff to player i over all those periods $0 \leq t' \leq t$ when he played action j, including some *initial propensity* $a_{ij}(0) > 0$. The cumulative payoff matching rule stipulates that in period $t + 1$, player i chooses action j with probability

$$p_{ij}(t+1) = a_{ij}(t) / \sum\nolimits_k a_{ik}(t). \qquad (1)$$

Notice that the distribution has full support given the assumption that the initial propensities are positive. This idea was first proposed by the psychologist Nathan Herrnstein (1970) to explain certain types of animal behaviour, and falls under the more general rubric of *reinforcement learning* (Bush and Mosteller 1951; Suppes and Atkinson 1960; Cross 1983). The key feature of a reinforcement model is that the probability of choosing an action increases monotonically with the total payoff it has generated in the past (on the assumption that the payoffs are positive). In other words, taking an action and receiving a positive payoff *reinforces* the tendency to take that same action again. This means, in particular, that play can become concentrated on certain actions simply because they were played early and often, that is, play can be *habit-forming* (Roth and Erev 1995; Erev and Roth 1998).

Reinforcement models differ in various details that materially affect their theoretical behaviour as well as their empirical plausibility. Under cumulative payoff matching, for example, the payoffs are not discounted, which means that current payoffs have an impact on current behaviour that diminishes as $1/t$. Laboratory experiments suggest, however, that recent payoffs matter more than those long past (Erev and Roth 1998); furthermore, the rate of discounting has implications for the asymptotic properties of such models (Arthur 1991).

Another variation in this class of models relies on the concept of an *aspiration level*. This is a level of payoffs, sometimes endogenously determined by past play, that triggers a change in a player's behaviour when current payoffs fall below the level and inertial behaviour when payoffs are above the level. The theoretical properties of these models have been studied for 2×2 games, but relatively little is known about their behaviour in general games (Börgers and Sarin 2000; Cho and Matsui 2005).

Next we turn to a class of adaptive heuristics based on the notion of minimizing *regret*, about which more is known in a theoretical sense. Fix a particular player and let $\alpha(t)$ denote the average per period payoff that she received over all periods $t' \leq t$. Let $\alpha_j(t)$ denote the average payoff she *would have* received by playing action j in every period through t, on the assumption that the opponents played as they actually did. The difference $r_j(t) = \alpha_j(t) - \alpha(t)$ is the subject's *unconditional regret* from not having played j in every period through t. (In the computer science literature this is known as *external regret*; see Greenwald and Gondek 2002.)

The following simple heuristic was proposed by Hart and Mas-Colell (2000, 2001) and is known as *unconditional regret matching*: play each action with a probability that is proportional to the positive part of its unconditional regret, that is,

$$p_j(t+1) = \left[r_j(t)\right]_+ / \sum\nolimits_k \left[r_k(t)\right]_+. \qquad (2)$$

This learning rule has the following remarkable property: when used by any one player, his regrets become non-positive almost surely as t goes to infinity *irrespective of the behaviour of the other players*. When all players use the rule, their time average behaviour converges almost surely to a generalization of correlated

equilibrium known as the *Hannan set* or the *coarse correlated equilibrium set* (Hannan 1957; Moulin and Vial 1978; Hart and Mas-Colell 2000; Young 2004).

In general, a *coarse correlated equilibrium* (CCE) is a probability distribution over outcomes (joint actions) such that, given a choice between (*a*) committing *ex ante* to whatever joint action will be realized, and (*b*) committing *ex ante* to a fixed action, given that the others are committed to playing their part of whatever joint action will be realized, every player weakly prefers the former option. By contrast, a *correlated equilibrium* (CE) is a distribution such that, after a player's part of the realized joint action has been disclosed, he would just as soon play it as something else, given that the others are going to play their part of the realized joint action. It is straightforward to show that the coarse correlated equilibria form a convex set that contains the set of correlated equilibria (Young 2004, ch. 3).

The heuristic specified in Eq. (2) belongs to a large family of rules whose time-average behaviour converges almost surely to the coarse correlated equilibrium set; equivalently, that assures no long-run regret for all players simultaneously. For example, this property holds if we let $p_j(t+1) = [r_j(t)]_+^\theta / \sum_k [r_k(t)]_+^\theta$ for some exponent $\theta > 0$; one may even take different exponents for different players. Notice that these heuristics put positive probability only on actions that would have done strictly better (on average) than the player's realized average payoff. These are sometimes called *better reply rules*. Fictitious play, by contrast, puts positive probability only on action (s) that would have done *best* against the opponents' frequency distribution of play.

Fictitious play does not necessarily converge to the coarse correlated equilibrium set (CCES); indeed, in some 2×2 coordination games fictitious play causes perpetual miscoordination, in which case both players have unconditional long-run regret (Fudenberg and Kreps 1993; Young 1993). By choosing θ to be very large, however, we see that there exist better reply rules that are arbitrarily close to fictitious play and that do converge almost surely to the CCES. Fudenberg and Levine (1995, 1998, 1999) and Hart and Mas-Colell (2001) give general conditions under which stochastic forms of fictitious play converge in time average to the CCES.

Without complicating the adjustment process too much, one can construct rules whose time average behaviour converges almost surely to the *correlated equilibrium set* (CES). To define this class of heuristics we need to introduce the notion of conditional regret. Given a history of play through time t and a player i, consider the change in per period payoff if i had played action k in all those periods $t' \leq t$ when he actually played action j (and the opponents played what they did). If the difference is positive, player i has conditional regret – he wishes he had played k instead of j. Formally, i's *conditional regret* at playing j instead of k up through time t, r^i_{jk}, is $1/t$ times the increase in payoff that would have resulted from playing k instead of j in all periods $t' \leq t$. Notice that the average is taken over all t periods to date; hence, if j was not played very often, r^i_{jk} will be small.

Consider the following *conditional regret matching* heuristic proposed by Hart and Mas-Colell (2000): if a given agent played action j in period t, then in period $t+1$ he plays according to the distribution

$$\begin{aligned} &qk(t+1) = \varepsilon r_{jk}(t)_+ \text{ for all } k \neq j, \\ &\text{and } q_j(t+1) = 1 - \varepsilon \sum_{k \neq j} r_{jk}(t)_+. \end{aligned} \tag{3}$$

In effect $1 - \varepsilon$ is the degree of inertia, which must be large enough that $q_k(t+1)$ 1Þ is non-negative for all realizations of the conditional regrets $r_{jk}(t)$. If all players use conditional regret matching and ε is sufficiently small, then almost surely the joint frequency of play converges to the set of correlated equilibria (Hart and Mas-Colell 2000). Notice that *pointwise* convergence is not guaranteed; the result says only that the empirical distribution converges to a convex *set*. In particular, the players' time-average behaviour may wander from one correlated equilibrium to another. It should also be remarked that, if a single player uses conditional regret matching, there is no

assurance that his conditional regrets will become non-positive over time unless we assume that the other players use the same rule. This stands in contrast to unconditional regret matching, which assures non-positive unconditional regret for any player who uses it irrespective of the behaviour of the other players. One can, however, design more sophisticated updating procedures that unilaterally assure no conditional regret; see for example Foster and Vohra (1999), Fudenberg and Levine (1998, ch. 4), Hart and Mas-Colell (2000), and Young (2004, ch. 4).

A natural question now arises: do there exist simple heuristics that allow the players to learn *Nash* equilibrium instead of correlated or still coarser forms of equilibrium? The answer depends on how demanding we are about the long-run convergence properties of the learning dynamic. Notice that the preceding results on regret matching were concerned solely with time-average behaviour; no claim was made that period-by-period behaviour converges to any notion of equilibrium. Yet surely it is period-by-period behaviour that is most relevant if we want to assert that the players have 'learned' to play equilibrium. It turns out that it is very difficult to design adaptive learning rules under which period-by-period behaviour converges almost surely to Nash equilibrium in any finite game, unless one builds in some form of coordination among the players (Hart and Mas-Colell 2003, 2006). The situation becomes even more problematic if one insists on fully rational, Bayesian learning. In this case it can be shown that there exist games of incomplete information in which no form of Bayesian rational learning causes period-by-period behaviours to come close to Nash equilibrium behaviour even in a probabilistic sense (Jordan 1991, 1993; Foster and Young 2001; Young 2004; see also learning and evolution in games: belief learning).

If one does not insist on full rationality, however, one can design stochastic adaptive heuristics that cause period-by-period behaviours to come close to Nash equilibrium – indeed close to subgame perfect equilibrium – most of the time (without necessarily *converging* to an equilibrium). Here is one approach due to Foster and Young (2003); for related work see Foster and Young (2006) and Germano and Lugosi (2007). Let G be a finite n-person game that is played infinitely often. At each point in time, each player thinks that the others are playing i.i.d. strategies. Specifically, at time t player i thinks that j is playing the i.i.d strategy $p_j(t)$ on j's action space, and that the opponents are playing independently; that is, their joint strategies are given by the product distribution $\mathrm{p}_{-i}(t) = \prod_{j \neq i} p_j(t)$. Suppose that i's best response is to play a smoothed best response to p_{-1} (t). Specifically, assume that i plays each action j with a probability proportional to $e^{\beta u_i(j, p_{-i})}$, where $u_i(j, p_{-i})$ is i's expected utility from playing j in every period when the opponents play p_{-i}, and $\beta > 0$ is a *response parameter*. This is known as a *quantal* or *log linear* response function. For brevity, denote i's response in period t by $q_i^{\beta}(t)$; this depends, of course, on p_{-i} (t). Player i views p_{-i} (t) as a hypothesis that he wishes to test against data. After first adopting this hypothesis he waits for a number of periods (say s) while he observes the opponents' behaviour, all the while playing $q_i^{\beta}(t)$. After s periods have elapsed, he compares the empirical frequency distribution of the opponents' play during these periods with his hypothesis. Notice that both the empirical frequency distribution and the hypothesized distribution lie in the same compact subset of Euclidean space. If the two differ by more than some tolerance level τ (in the Euclidean metric), he rejects his current hypothesis and chooses a new one.

In choosing a new hypothesis, he may wish to take account of information revealed during the course of play, but we shall also assume he engages in some *experimentation*. Specifically, let us suppose that he chooses a new hypothesis according to a probability density that is uniformly bounded away from zero on the space of hypotheses. One can show the following: given any $\varepsilon > 0$, if the response parameter β is sufficiently large, the test tolerance τ is sufficiently small (given β), and the amount of data collected s is sufficiently large (given β and τ), then the players' *period-by-period* behaviours constitute an ε-equilibrium of the stage game G at least $1 -$

ε of the time (Foster and Young 2003). In other words, classical statistical hypothesis testing is a heuristic for learning Nash equilibria of the stage game. Moreover, if the players adopt hypotheses that condition on history, they can learn complex equilibria of the repeated game, including forms of subgame perfect equilibrium.

The theoretical literature on strategic learning has advanced rapidly in recent years. A much richer class of learning models has been identified since the mid-1990s, and more is known about their long-run convergence properties. There is also a greater understanding of the various kinds of equilibrium that different forms of learning deliver. An important open question is how these theoretical proposals relate to the empirical behaviour of laboratory subjects. While there is no reason to think that any of these rules can fully explain subjects' behaviour, they can nevertheless play a useful role by identifying phenomena that experimentalists should look for. In particular, the preceding discussion suggests that weaker forms of equilibrium may turn out to be more robust predictors of long-run behaviour than is Nash equilibrium.

See Also

- ▶ Behavioural Game Theory
- ▶ Learning and Evolution in Games: Belief Learning

Bibliography

Arthur, W.B. 1991. Designing agents that act like human agents: A behavioral approach to bounded rationality. *American Economic Association, Papers and Proceedings* 81: 353–359.

Börgers, T., and R. Sarin. 2000. Naïve reinforcement learning with endogenous aspirations. *International Economic Review* 31: 921–950.

Bush, R.R., and F. Mosteller. 1951. A mathematical model for simple learning. *Psychological Review* 58: 313–323.

Cho, I.-K., and A. Matsui. 2005. Learning aspiration in repeated games. *Journal of Economic Theory* 124: 171–201.

Cross, J. 1983. *A theory of adaptive economic behavior*. Cambridge: Cambridge University Press.

Erev, I., and A.E. Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88: 848–881.

Foster, D.P., and R. Vohra. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29: 7–35.

Foster, D.P., and H.P. Young. 2001. On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences of the United States of America* 98(222): 12848–12853.

Foster, D.P., and H.P. Young. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45: 73–96.

Foster, D.P., and H.P. Young. 2006. Regret testing: Learning Nash equilibrium without knowing you have an opponent. *Theoretical Economics* 1: 341–367.

Fudenberg, D., and D. Kreps. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5: 320–367.

Fudenberg, D., and D. Levine. 1995. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19: 1065–1090.

Fudenberg, D., and D. Levine. 1998. *The theory of learning in games*. Cambridge, MA: MIT Press.

Fudenberg, D., and D. Levine. 1999. Conditional universal consistency. *Games and Economic Behavior* 29: 104–130.

Germano, F., and G. Lugosi. 2007. Global Nash convergence of Foster and Young's regret testing. *Games and Economic Behavior* 60: 135–154.

Greenwald, A., and D. Gondek. 2002. On no-regret learning and game-theoretic equilibria. *Journal of Machine Learning* 1: 1–20.

Hannan, J. 1957. Approximation to Bayes risk in repeated plays. In *Contributions to the theory of games*, ed. M. Dresher, A.W. Tucker, and P. Wolfe, Vol. 3. Princeton: Princeton University Press.

Hart, S., and A. Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68: 1127–1150.

Hart, S., and A. Mas-Colell. 2001. A general class of adaptive strategies. *Journal of Economic Theory* 98: 26–54.

Hart, S., and A. Mas-Colell. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93: 1830–1836.

Hart, S., and A. Mas-Colell. 2006. Stochastic uncoupled dynamics and Nash equilibrium. *Games and Economic Behavior* 57: 286–303.

Herrnstein, R.J. 1970. On the law of effect. *Journal of the Experimental Analysis of Behavior* 13: 243–266.

Jordan, J.S. 1991. Bayesian learning in normal form games. *Games and Economic Behavior* 3: 60–81.

Jordan, J.S. 1993. Three problems in learning mixed-strategy equilibria. *Games and Economic Behavior* 5: 368–386.

Moulin, H., and J.P. Vial. 1978. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Games Theory* 7: 201–221.

Roth, A.E., and I. Erev. 1995. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8: 164–212.
Suppes, P., and R. Atkinson. 1960. *Markov learning models for multiperson interaction*. Stanford: Stanford University Press.
Young, H.P. 1993. The evolution of conventions. *Econometrica* 61: 57–84.
Young, H.P. 2004. *Strategic learning and its limits*. Oxford: Oxford University Press.

Learning and Evolution in Games: An Overview

William H. Sandholm

Abstract

We provide a taxonomy and brief overview of the theory of learning and evolution in games.

Keywords

Adaptive learning; Belief learning; Convergence; Deterministic evolutionary dynamics; Dirichlet distributions; Evolutionarily stable strategies; Evolutionary game theory; Fictitious play; Forward induction; Learning and evolution in games; Markov processes; Mathematical biology; Natural selection; Purification; Rational learning; Regret; Regret matching; Repeated games; Replicator dynamic; Revision protocols; Risk dominant equilibrium; Stochastic evolutionary dynamics; Subgame perfection

JEL Classifications

C7

The author thanks John Nachbar for a number of helpful conversations and for sharing his expertise on coordinated Bayesian learning. Financial support under NSF Grants SES-0092145 and SES-0617753 is gratefully acknowledged.

The theory of learning and evolution in games provides models of disequilibrium behaviour in strategic settings. Much of the theory focuses on whether and when disequilibrium behaviour will resolve in equilibrium play, and, if it does, on predicting which equilibrium will be played. But the theory also offers techniques for characterizing perpetual disequilibrium play.

A Taxonomy

Models from *evolutionary game theory* consider the behaviour of large populations in strategic environments. In the biological strand of the theory, agents are genetically programmed to play fixed actions, and changes in the population's composition are the result of natural selection and random mutations. In economic approaches to the theory, agents actively choose which actions to play using simple myopic rules, so that changes in aggregate behaviour are the end result of many individual decisions. *Deterministic evolutionary dynamics*, usually taking the form of ordinary differential equations, are used to describe behaviour over moderate time spans, while *stochastic evolutionary dynamics*, modelled using Markov processes, are more commonly employed to study behaviour over very long time spans.

Models of *learning in games* focus on the behaviour of small groups of players, one of whom fills each role in a repeated game. These models too can be partitioned into two categories. Models of *heuristic learning* (or *adaptive learning*) resemble evolutionary models, in that their players base their decisions on simple myopic rules. One sometimes can distinguish the two sorts of models by the inputs to the agents' decision rules. In both the stochastic evolutionary model of c, Kandori, Mailath and Rob (1993) and the heuristic learning model of Young (1993), agents' decisions take the form of noisy best responses. But in the former model agents evaluate each action by its performance against the population's current behaviour, while in the latter they consider performance against the time averages of opponents' past play.

In models of *coordinated Bayesian learning* (or *rational learning*), each player forms explicit

beliefs about the repeated game strategies employed by other players, and plays a best response to those beliefs in each period. The latter models assume a degree of coordination of players' prior beliefs that is sufficient to ensure that play converges to Nash equilibrium. By dropping this coordination assumption, one obtains the more general class of *Bayesian learning* (or *belief learning*) models. Since such models can entail quite naive beliefs, belief learning models overlap with heuristic learning models – see section "Learning in Games" below.

Evolutionary Game Theory

The roots of evolutionary game theory lie in mathematical biology. Maynard Smith and Price (1973) introduced the equilibrium notion of an *evolutionarily stable strategy* (or ESS) to capture the possible stable outcomes of a dynamic evolutionary process by way of a static definition. Later, Taylor and Jonker (1978) offered the *replicator dynamic* as an explicitly dynamic model of the natural selection process. The decade that followed saw an explosion of research on the replicator dynamic and related models of animal behaviour, population ecology, and population genetics: see Hofbauer and Sigmund (1988).

In economics, evolutionary game theory studies the behaviour of populations of strategically interacting agents who actively choose among the actions available to them. Agents decide when to switch actions and which action to choose next using simple myopic rules known as *revision protocols* (see Sandholm 2006). A population of agents, a game, and a revision protocol together define a stochastic process – in particular, a Markov process – on the set of population states.

Deterministic Evolutionary Dynamics

How the analysis proceeds depends on the time horizon of interest. Suppose that for the application in question, our interest is in moderate time spans. Then if the population size is large enough, the idiosyncratic noise in agent's choices is averaged away, so that the evolution of aggregate behaviour follows an almost deterministic path (Benaïm and Weibull 2003). This path is described by a solution to an ordinary differential equation. For example, Björnerstedt and Weibull (1996) and Schlag (1998) show that if agents use certain revision protocols based on imitation of successful opponents, then the population's aggregate behaviour follows a solution to Taylor and Jonker's (1978) replicator dynamic. This argument provides an alternative, economic interpretation of this fundamental evolutionary model.

Much of the literature on deterministic evolutionary dynamics focuses on connections with traditional game theoretic solution concepts. For instance, under a wide range of deterministic dynamics, all Nash equilibria of the underlying game are rest points. While some dynamics (including the replicator dynamic) have additional non-Nash rest points, there are others under which rest points and Nash equilibria are identical (Brown and von Neumann, 1950; Smith 1984; Sandholm 2006).

A more important question, though, is whether Nash equilibrium will be approached from arbitrary disequilibrium states. For certain specific classes of games, general convergence results can be established (Hofbauer 2000; Sandholm 2007). But beyond these classes, convergence cannot be guaranteed. One can construct games under which no reasonable deterministic evolutionary dynamic will converge to equilibrium – instead, the population cycles through a range of disequilibrium states forever (Hofbauer and Swinkels 1996; Hart and Mas-Colell 2003). More surprisingly, one can construct games in which nearly all deterministic evolutionary dynamics not only cycle for ever, but also fail to eliminate strictly dominated strategies (Hofbauer and Sandholm 2006). If we truly are interested in modelling the dynamics of behaviour, these results reveal that our predictions cannot always be confined to equilibria; rather, more complicated limit phenomena like cycles and chaotic attractors must also be permitted as predictions of play.

Stochastic Evolutionary Dynamics

If we are interested in behaviour over very long time horizons, deterministic approximations are no longer valid, and we must study our original

Markov process directly. Under certain non-degeneracy assumptions, the long-run behaviour of this process is captured by its unique stationary distribution, which describes the proportion of time the process spends in each population state.

While stochastic evolutionary processes can be more difficult to analyse than their deterministic counterparts, they also permit us to make surprisingly tight predictions. By making the amount of noise in agents' choice rules vanishingly small, one can often ensure that all mass in the limiting stationary distribution is placed on a single population state. This *stochastically stable state* provides a unique prediction of play even in games with multiple strict equilibria (Foster and Young 1990; Kandori, Mailath and Rob, 1993).

The most thoroughly studied model of stochastic evolution considers agents who usually play a best response to the current population state, but who occasionally choose a strategy at random. Kandori, Mailath and Rob (1993) show that if the agents are randomly matched to play a symmetric 2×2 coordination game, then taking the probability of 'mutations' to zero generates a unique stochastically stable state. In this state, called the *risk dominant equilibrium*, all agents play the action that is optimal against an opponent who is equally likely to choose each action.

Selection results of this sort have since been extended to cases in which the underlying game has an arbitrary number of strategies, as well as to settings in which agents are positioned on a fixed network, interacting only with neighbours (see Kandori and Rob 1995; Blume 2003; Ellison 1993; 2000). Stochastic stability has also been employed in contexts where the underlying game has a nontrivial extensive form; these analyses have provided support for notions of backward induction (for example, subgame perfection) and forward induction (for example, signalling game equilibrium refinements): see Nöldeke and Samuelson (1993) and Hart (2002).

Still, these selection results must be interpreted with care. When the number of agents is large or the rate of 'mutation' is small, states that fail to be stochastically stable can be coordinated upon for great lengths of time (Binmore, Samuelson and Vaughan, 1995). Consequently, if the relevant time span for the application at hand is not long enough, the stochastically stable state may not be the only reasonable prediction of behaviour.

Learning in Games

Heuristic Learning

Learning models study disequilibrium adjustment processes in repeated games. Like evolutionary models, heuristic learning models assume that players employ simple myopic rules in deciding how to act. In the simplest of these models, each player decides how to act by considering the payoffs he has earned in the past. For instance, under reinforcement learning (Börgers and Sarin 1997; Erev and Roth 1998), agents choose each strategy with probability proportional to the total payoff that the strategy has earned in past periods.

By considering rules that look not only at payoffs earned, but also at payoffs foregone, one can obtain surprisingly strong convergence results. Define a player's *regret* for (not having played) action a to be the difference between the average payoff he would have earned had he always played a in the past, and the average payoff he actually received. Under *regret matchingt*, each action whose regret is positive is chosen with probability proportional to its regret. Hart and Mas-Colell (2000) show that regret matching is a *consistent* repeated game strategy: it forces a player's regret for each action to become nonpositive. If used by all players, regret matching ensures that their time-averaged behaviour converges to the set of *coarse correlated equilibria* of the underlying game. (*Coarse correlated equilibrium* is a generalization of correlated equilibrium under which players' incentive constraints must be satisfied at the *ex ante* stage rather than at the interim stage: see Young 2004.)

Some of the most striking convergence results in the evolution and learning literature establish a stronger conclusion: namely, convergence of time-averaged behaviour to the set of *correlated equilibria*, regardless of the game at hand. The original result of this sort is due to Foster and Vohra (1997; 1998), who prove the result by

constructing a calibrated procedure for forecasting opponents' play. A *forecasting procedure* produces probabilistic forecasts of how opponents will act. The procedure is *calibrated* if in those periods in which the forecast is given by the probability vector p, the empirical distribution of opponents' play is approximately p. It is not difficult to show that if players always choose myopic best responses to calibrated forecasts, then their time-averaged behaviour converges to the set of correlated equilibria.

Hart and Mas-Colell (2000) construct simpler procedures – in particular, procedures that define conditionally consistent repeated game strategies – also ensure convergence to correlated equilibrium. A repeated game strategy is *conditionally consistent* if for each frequently played action a, the agent would not have been better off had he always played an alternative action a' in place of a. As a matter of definition, the use of conditionally consistent strategies by all players leads time-averaged behavior to converge to the set of correlated equilibria.

Another variety of heuristic learning models, based on *random search and independent verification*, ensures a stochastic form of convergence to Nash equilibrium regardless of the game being played (Foster and Young 2003). However, in these models the time required before equilibrium is first reached is quite long, making them most relevant to applications with especially long time horizons.

In some heuristic learning models, players use simple rules to predict how opponents will behave, and then respond optimally to those predictions. The leading examples of such models are *fictitious play* and its stochastic variants (Brown 1951; Fudenberg and Kreps 1993): in these models, the prediction about an opponents' next period play is given by the empirical frequencies of his past plays. Beginning with Robinson (1951), many authors have proved convergence results for standard and stochastic fictitious play in specific classes of games (see Hofbauer and Sandholm (2002) for an overview). But as Shapley (1964) and others have shown, these models do not lead to equilibrium play in all games.

Coordinated Bayesian Learning

The prediction rule underlying two-player fictitious play can be described by a belief about the opponent's repeated game strategy that is updated using Bayes's rule in the face of observed play. This belief specifies that the opponent choose his stage game actions in an i.i.d. fashion, conditional on the value of an unknown parameter. (In fact, the player's beliefs about this parameter must come from the family of Dirichlet distributions, the conjugate family of distributions for multinomial trials.) Evidently, each player's beliefs about his opponent are wrong: player 1 believes that player 2 chooses actions in an i.i.d. fashion, whereas player 2 actually plays optimally in response to his own (i.i.d.) predictions about player 1's behaviour. It is therefore not surprising that fictitious play processes do not converge in all games.

In models of *coordinated Bayesian learning* (or *rational learning*), it is not only supposed that players form and respond optimally to beliefs about the opponent's repeated game strategy; it is also assumed that the players' initial beliefs are coordinated in some way. The most studied case is one in which prior beliefs satisfy an absolute continuity condition: if the distribution over play paths generated by the players' actual strategies assigns positive probability to some set of play paths, then so must the distribution generated by each player's prior. A strong sufficient condition for absolute continuity is that each player's prior assigns a positive probability to his opponent's actual strategy.

The fundamental result in this literature, due to Kalai and Lehrer (1993), shows that under absolute continuity, each player's forecast along the path of play is asymptotically correct, and the path of play is asymptotically consistent with Nash equilibrium play in the repeated game. Related convergence results have been proved for more complicated environments in which each player's stage game payoffs are private information (Jordan 1995; Nyarko 1998). If the distributions of players types are continuous, then the sense in which play converges to equilibrium can involve a form of purification: while actual play is pure, it appears random to an outside observer.

How much coordination of prior beliefs is needed to prove convergence to equilibrium play? Nachbar (2005) proves that for a large class of repeated games, for any belief learning model, there are no prior beliefs that satisfy three criteria: learnability, consistency with optimal play, and diversity. Thus, if players can learn to predict one another's behaviour, and are capable of responding optimally to their updated beliefs, then each player's beliefs about his opponents must rule out some seemingly natural strategies a priori. In this sense, the assumption of coordinated prior beliefs that ensures convergence to equilibrium in rational learning models does not seem dramatically weaker than a direct assumption of equilibrium play.

For additional details about the theory of learning and evolution in games, we refer the reader to the entries on specific topics listed in the cross-references below.

See Also

- ▶ Deterministic Evolutionary Dynamics
- ▶ Learning and Evolution in Games: Adaptive Heuristics
- ▶ Learning and Evolution in Games: Belief Learning
- ▶ Learning and Evolution in Games: ESS
- ▶ Stochastic Adaptive Dynamics

Bibliography

Benaïm, M., and J.W. Weibull. 2003. Deterministic approximation of stochastic evolution in games. *Econometrica* 71: 873–903.

Binmore, K., L. Samuelson, and R. Vaughan. 1995. Musical chairs: Modeling noisy evolution. *Games and Economic Behavior* 11: 1–35.

Björnerstedt, J., and J.W. Weibull. 1996. Nash equilibrium and evolution by imitation. In *The rational foundations of economic behavior*, ed. K.J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt. New York: St. Martin's Press.

Blume, L.E. 2003. How noise matters. *Games and Economic Behavior* 44: 251–271.

Börgers, T., and R. Sarin. 1997. Learning through reinforcement and the replicator dynamics. *Journal of Economic Theory* 77: 1–14.

Brown, G.W. 1951. Iterative solutions of games by fictitious play. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.

Brown, G.W., and J. von Neumann. 1950. Solutions of games by differential equations. In *Contributions to the theory of games I*, Annals of mathematics studies, vol. 24, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61: 1047–1071.

Ellison, G. 2000. Basins of attraction, long run equilibria, and the speed of step-by-step evolution. *Review of Economic Studies* 67: 17–45.

Erev, I., and A.E. Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88: 848–881.

Foster, D.P., and R. Vohra. 1997. Calibrated learning and correlated equilibrium. *Games and Economic Behavior* 21: 40–55.

Foster, D.P., and R. Vohra. 1998. Asymptotic calibration. *Biometrika* 85: 379–390.

Foster, D.P., and H.P. Young. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38: 219–232.

Foster, D.P., and H.P. Young. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45: 73–96.

Fudenberg, D., and D.M. Kreps. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5: 320–367.

Fudenberg, D., and D.K. Levine. 1998. *Theory of learning in games*. Cambridge: MIT Press.

Hart, S. 2002. Evolutionary dynamics and backward induction. *Games and Economic Behavior* 41: 227–264.

Hart, S., and A. Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68: 1127–1150.

Hart, S., and A. Mas-Colell. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93: 1830–1836.

Hofbauer, J. 2000. From Nash and Brown to Maynard Smith: Equilibria, dynamics, and ESS. *Selection* 1: 81–88.

Hofbauer, J., and W.H. Sandholm. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70: 2265–2294.

Hofbauer, J., and W.H. Sandholm. 2006. *Survival of dominated strategies under evolutionary dynamics*. Working paper. University of Vienna and University of Wisconsin.

Hofbauer, J., and K. Sigmund. 1988. *Theory of evolution and dynamical systems*. Cambridge: Cambridge University Press.

Hofbauer, J., and J.M. Swinkels. 1996. *A universal Shapley example*. Working paper. University of Vienna and Northwestern University.

Jordan, J.S. 1995. Bayesian learning in repeated games. *Games and Economic Behavior* 9: 8–20.

Kalai, E., and E. Lehrer. 1993. Rational learning leads to Nash equilibrium. *Econometrica* 61: 1019–1045.

Kandori, M., G.J. Mailath, and R. Rob. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61: 29–56.
Kandori, M., and R. Rob. 1995. Evolution of equilibria in the long run: A general theory and applications. *Journal of Economic Theory* 65: 383–414.
Maynard Smith, J., and G.R. Price. 1973. The logic of animal conflict. *Nature* 246: 15–18.
Nachbar, J.H. 2005. Beliefs in repeated games. *Econometrica* 73: 459–480.
Nöldeke, G., and L. Samuelson. 1993. An evolutionary analysis of backward and forward induction. *Games and Economic Behavior* 5: 425–454.
Nyarko, Y. 1998. Bayesian learning and convergence to Nash equilibria without common priors. *Economic Theory* 11: 643–655.
Robinson, J. 1951. An iterative method of solving a game. *Annals of Mathematics* 54: 296–301.
Sandholm, W.H. 2006. *Pairwise comparison dynamics and evolutionary foundations for Nash equilibrium*. Working paper. University of Wisconsin.
Sandholm, W.H. 2007. *Population games and evolutionary dynamics*. Cambridge, MA: MIT Press.
Schlag, K.H. 1998. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory* 78: 130–156.
Shapley, L.S. 1964. Some topics in two person games. In *Advances in game theory*, Annals of mathematics studies, vol. 52, ed. M. Dresher, L.S. Shapley, and A.W. Tucker. Princeton: Princeton University Press.
Smith, M.J. 1984. The stability of a dynamic model of traffic assignment: An application of a method of Lyapunov. *Transportation Science* 18: 245–252.
Taylor, P.D., and L. Jonker. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40: 145–156.
Young, H.P. 1993. The evolution of conventions. *Econometrica* 61: 57–84.
Young, H.P. 2004. *Strategic learning and its limits*. Oxford: Oxford University Press.

Learning and Evolution in Games: Belief Learning

John Nachbar

Abstract

In the context of learning in games, *belief learning* refers to models in which players are engaged in a dynamic game and each player optimizes with respect to a *prediction rule* that gives a forecast of next-period opponent behaviour as a function of the current history. This article focuses on the most studied class of dynamic games, namely, two-player discounted repeated games with finite stage game action sets and perfect monitoring.

Keywords

Belief learning; Best-response dynamics; Convergence; Cournot, A. A.; Distributional strategies; Dynamic games; Fictitious play; Folk theorem; Learnable best-response property; Mixed strategy equilibrium; Myopia; Perfect monitoring; Prediction rules; Purification; Repeated games

JEL Classifications

C7

In the context of learning in games, *belief learning* refers to models in which players are engaged in a dynamic game and each player optimizes, or ε optimizes, with respect to a *prediction rule* that gives a forecast of next period opponent behaviour as a function of the current history. This article focuses on the most studied class of dynamic games, two-player discounted repeated games with finite stage game action sets and perfect monitoring. An important example of a dynamic game that violates perfect monitoring and therefore falls outside this framework is Fudenberg and Levine (1993). For a more comprehensive survey of belief learning, see Fudenberg and Levine (1998).

The earliest example of belief learning is the *best-response dynamics* of Cournot (1838). In Cournot's model, each player predicts that her opponent will repeat next period whatever action her opponent chose in the previous period.

The most studied belief learning model is *fictitious play* (Brown, 1951), and its variants. In fictitious play, each player predicts that the probability that her opponent will play an action, say L, next period is a weighted sum of an initial probability on L and the frequency with which L has been chosen to date. The weight on the frequency is $t/(t+k)$, where t is the number of periods thus far and $k > 0$ is a parameter. The larger is k, the more

periods for which the initial probability significantly affects forecasting.

The remainder of this article discusses four topics: (1) belief learning versus Bayesian learning, (2) convergence to equilibrium, (3) special issues in games with payoff uncertainty, and (4) sensible beliefs.

Belief Learning Versus Bayesian Learning

Recall that, in a repeated game, a behaviour strategy gives, for every history, a probability over the player's stage game actions next period. In a Bayesian model, each player chooses a behaviour strategy that best responds to a *belief*, a probability distribution over the opponent's behaviour strategies.

Player 1's prediction rule about player 2 is mathematically identical to a behaviour strategy for player 2. Thus, any belief learning model is equivalent to a Bayesian model in which each player optimizes with respect to a belief that places probability 1 on her prediction rule, now reinterpreted as the opponent's behaviour strategy.

Conversely, any Bayesian model is equivalent to a belief learning model. Explicitly, for any belief over player 2's behaviour strategies there is a degenerate belief, assigning probability 1 to a particular behaviour strategy, that is equivalent in the sense that both beliefs induce the same distributions over play in the game, no matter what behaviour strategy player 1 herself adopts. This is a form of Kuhn's theorem (Kuhn, 1964). I refer to the behaviour strategy used in the degenerate belief as a *reduced form* of the original belief. Thus, any Bayesian model is equivalent to a Bayesian model in which each player's belief places probability 1 on the reduced form, and any such Bayesian model is equivalent to a belief learning model.

As an example, consider fictitious play. I focus on stage games with just two actions, L and R. By an i.i.d. strategy for player 2, I mean a behaviour strategy in which player 2 plays L with probability q, independent of history. Thus, if $q = 1/2$, then player 2 always randomizes 50:50 between L and R. Fictitious play is equivalent to a degenerate Bayesian model in which each player places probability 1 on the fictitious play prediction rule, and one can show that this is equivalent in turn to a non-degenerate Bayesian model in which the belief is represented as a beta distribution over q. The uniform distribution over q, for example, corresponds to taking the initial probability of L to be 1/2 and the parameter k to be 2.

There is a related but distinct literature in which players optimize with respect to *stochastic* prediction rules. In some cases (for example, Foster and Young, 2003), these models have a quasi-Bayesian interpretation: most of the time, players optimize with respect to fixed prediction rules, as in a Bayesian model, but occasionally players switch to new prediction rules, implicitly abandoning their priors.

Convergence to Equilibrium

Within the belief learning literature, the investigation of convergence to equilibrium play splits into two branches. One branch investigates convergence within the context of specific classes of belief learning models. The best-response dynamics, for example, converge to equilibrium if the stage game is solvable by the iterated deletion of strictly dominated strategies. See Bernheim (1984) and, for a more general class of models, Milgrom and Roberts (1991). For an ε optimizing variant of fictitious play, convergence to approximate equilibrium play obtains for all zero-sum games, all games with an interior ESS, and all common interest games, in addition to all games that are strict dominance solvable, with the approximation closer the smaller is ε. Somewhat weaker convergence results are available for supermodular games. These claims follow from results in Hofbauer and Sandholm (2002).

In the results surveyed above, convergence is to repeated play of a single-stage game Nash equilibrium; in the case of ε fictitious play, this equilibrium may be mixed. There is a large body of work on convergence that is weaker than what I am considering here. In particular, there has been

much work on convergence of the empirical marginal or joint distributions. For mixed strategy equilibrium, it is possible for empirical distributions to converge to equilibrium even though play does not resemble repeated equilibrium play; play may exhibit obvious cycles, for example. The study of convergence to equilibrium play is relatively recent and was catalysed by Fudenberg and Kreps (1993).

There are classes of games that cause convergence problems for many standard belief learning models, even when one considers only weak forms of convergence, such as convergence of the empirical marginal distributions (see Shapley, 1962; Jordan, 1993). Hart and Mas-Colell (2003, 2006) (hereafter HM) shed light on non-convergence by investigating learning models, including but not limited to belief learning models, that are *decoupled*, meaning that player 1's behaviour does not depend directly on player 2's stage game payoffs. A continuous time version of fictitious play fits into the framework of Hart and Mas-Colell (2003). The HM results imply that universal convergence is impossible for large classes of decoupled belief learning models: for any such model there exist stage games and initial conditions for which play fails to converge to equilibrium play.

The second branch of the literature, for which Kalai and Lehrer (1993a) (hereafter KL) is the central paper, takes a Bayesian perspective and asks what conditions on beliefs are sufficient to give convergence to equilibrium play. I find it helpful to characterize this literature in the following way. Say that a belief profile (giving a belief for each player) has the *learnable best-response property* (LBR) if there is a profile of best-response strategies (LBR strategies) such that, if the LBR strategies are played, then each player learns to predict the play path.

A player *learns to predict the play path* if her prediction of next period's play is asymptotically as good as if she knew her opponent's behaviour strategy. If the behaviour strategies call for randomization then players accurately predict the distribution over next period's play rather than the realization of next period's play. For example, consider a 2×2 game in which player 1 has stage game actions T and B and player 2 has stage game actions L and R. If player 2 is randomizing 50:50 every period and player 1 learns to predict the path of play, then for every ε there is a time, which depends on the realization of player 2's strategy, after which player 1's next period forecast puts the probability of L within ε of 1/2. (This statement applies to a set of play paths that arises with probability 1 with respect to the underlying probability model; I gloss over this sort of complication both here and below.) For a more complicated example, suppose that in period t player 2 plays L with probability $1 - \alpha$, where α is the frequency that the players have played the profile (B, R). If player 1 learns to predict the play path, then for any ε there is a time, which now depends on the realization of both players' strategies, after which player 1's next period forecast puts the probability of L within ε of $1 - \alpha$.

Naively, if LBR holds, and players are using their LBR strategies, then, in the continuation game, players are optimizing with respect to posterior beliefs that are asymptotically correct and so continuation behaviour strategies should asymptotically be in equilibrium. This intuition is broadly correct, but there are three qualifications.

First, in general, convergence is to Nash equilibrium play in the *repeated* game, not necessarily to repeated play of a single stage game equilibrium. If players are myopic (meaning that players optimize each period as though their discount factors were zero), then the set of equilibrium play paths comprises all possible sequences of stage game Nash equilibria, which is a very large set if the stage game has more than one equilibrium. If players are patient, then the folk theorem applies and the set of possible equilibrium paths is typically even larger.

Second, convergence is to an equilibrium play path, not necessarily to an equilibrium of the repeated game. The issue is that LBR implies accurate forecasting only along the play path. A player's predictions about how her opponent would respond to deviations may be grossly in error, for ever. Therefore, posterior beliefs need *not* be asymptotically correct and, unless players are myopic, continuation behaviour strategies need *not* be asymptotically in equilibrium. Kalai

and Lehrer (1993b) shows that behaviour strategies can be doctored at information sets off the play path so that the modified behaviour strategies are asymptotically in equilibrium yet still generate the same play path. This implies that the play path of the original strategy profile was asymptotically an equilibrium play path.

Third, the exact sense in which play converges to equilibrium play depends on the strength of learning. See KL and also Sandroni (1998).

KL shows that a strong form of LBR holds if beliefs satisfy an absolute continuity condition: each player assigns positive probability to any (measurable) set of play paths that has positive probability given the players' actual strategies. A sufficient condition for this is that each player assigns positive, even if extremely low, probability to her opponent's actual strategy, a condition that KL call *grain of truth*. Nyarko (1998) provides the appropriate generalization of absolute continuity for games with type space structures, including the games with payoff uncertainty discussed below.

Games with Payoff Uncertainty

Suppose that, at the start of the repeated game, each player is privately informed of his or her stage game payoff function, which remains fixed throughout the course of the repeated game. Refer to player i's stage game payoff function as her *payoff type*. Assume that the joint distribution over payoff functions is independent (to avoid correlation issues that are not central to my discussion) and commonly known.

Each player can condition her behaviour strategy in the repeated game on her realized payoff type. A mathematically correct way of representing this conditioning is via distributional strategies (see Milgrom and Weber, 1985).

For any belief about player 2, now a probability distribution over player 2's distributional strategies, and given the probability distribution over player 2's payoff types, there is a behaviour strategy for player 2 in the repeated game that is equivalent in the sense that it generates the same distribution over play paths. Again, this is essentially Kuhn's theorem. And again, I refer to this behaviour strategy as a *reduced form*.

Say that a player *learns to predict the play path* if her forecast of next period's play is asymptotically as good as if she knew the reduced form of her opponent's distributional strategy. This definition specializes to the previous one if the distribution over types is degenerate. If distributional strategies are in equilibrium then, in effect, each player is optimizing with respect to a degenerate belief that puts probability one on her opponent's actual distributional strategy and in this case players trivially learn to predict the path of play.

One can define LBR for distributional strategies and, as in the payoff certainty case, one can show that LBR implies convergence to equilibrium play in the repeated game with payoff types. More interestingly, there is a sense in which play converges to equilibrium play of the *realized* repeated game – the repeated game determined by the realized type profile. The central paper is Jordan (1991). Other important papers include KL (cited above), Jordan (1995), Nyarko (1998), and Jackson and Kalai (1999) (which studies recurring rather than repeated games).

Suppose first that the realized type profile has positive probability. In this case, if a player learns to predict the play path, then, as shown by KL, her forecast is asymptotically as good as if she knew both her opponent's distributional strategy *and* her opponent's realized type. LBR then implies that actual play, meaning the distribution over play paths generated by the realized behaviour strategies, converges to equilibrium play of the realized repeated game. For example, suppose that the type profile for matching pennies gets positive probability. In the unique equilibrium of repeated matching pennies, players randomize 50:50 in every period. Therefore, LBR implies that, if the matching pennies type profile is realized, then each player's behaviour strategy in the realized repeated game involves 50:50 randomization asymptotically.

If the distribution over types admits a continuous density, so that no type profile receives positive probability, then the form of convergence is more subtle. Suppose that players are myopic and that the realized stage game is like matching

pennies, with a unique and fully mixed equilibrium. Given myopia, the unique equilibrium of the realized repeated game calls for repeated play of the stage game equilibrium. In particular, it calls for players to randomize. It is not hard to show, however, that in a type space game with a continuous density, optimization calls for each player to play a pure strategy for almost every realized type. Thus, for almost every realized type profile in a neighbourhood of a game like matching pennies, actual play (again meaning the distribution over play paths generated by the realized behaviour strategies) cannot converge to equilibrium play, *even if the distributional strategies are in equilibrium*. Foster and Young (2001) provides a generalization for non-myopic players.

There is, however, a weaker sense in which play nevertheless does converge to equilibrium play in the realized repeated game. For simplicity, assume that each player knows the other's distributional strategy and that these strategies are in equilibrium. One can show that to an outsider observed play looks asymptotically like equilibrium play in the realized repeated game. In particular, if the realized game is like repeated matching pennies then observed play looks random. Moreover, to a player in the game, opponent behaviour looks random because, even though she knows her opponent's distributional strategy, she does not know her opponent's type. As play proceeds, each player in effect learns more about her opponent's type, but never enough to zero in on her opponent's realized, pure, behaviour strategy. Thus, when the distribution over types admits a continuous density, convergence to equilibrium involves a form of purification in the sense of Harsanyi (1973), a point that has been emphasized by Nyarko (1998) and Jackson and Kalai (1999).

Sensible Beliefs

A number of papers investigate classes of prediction rules that are sensible in that they exhibit desirable properties, such as the ability to detect certain kinds of patterns in opponent behaviour (see Aoyagi, 1996; Fudenberg and Levine, 1995, 1999; Sandroni, 2000).

Nachbar (2005) instead studies the issue of sensible beliefs from a Bayesian perspective. For simplicity, focus on learning models with known payoffs. Fix a belief profile, fix a subset of behaviour strategies for each player, and consider the following criteria for these subsets.

- *Learnability* – given beliefs, if players play a strategy profile drawn from these subsets then they learn to predict the play path.
- *Richness*. Informally (the formal statement is tedious), richness requires that if a behaviour strategy is included in one of the strategy subsets then certain variations on that strategy must be included as well. Richness, called CSP in Nachbar (2005), is satisfied automatically if the strategy subsets consist of all strategies satisfying a standard complexity bound, the same bound for both players. Thus richness holds if the subsets consist of all strategies with k-period memory, or all strategies that are automaton implementable, or all strategies that are Turing implementable, and so on.
- *Consistency* – each player's subset contains a best response to her belief.

The motivating idea is that, if beliefs are probability distributions over strategy subsets satisfying learnability, richness, and consistency, then beliefs are sensible, or at least are candidates for being considered sensible. Nachbar (2005) studies whether any such beliefs exist.

Consider, for example, the Bayesian interpretation of fictitious play in which beliefs are probability distributions over the i.i.d. strategies. The set of i.i.d. strategies satisfies learnability and richness. But for any stage game in which neither player has a weakly dominant action, the i.i.d. strategies violate consistency: any player who is optimizing will not be playing i.i.d.

Nachbar (2005) shows that this feature of Bayesian fictitious play extends to all Bayesian learning models. For large classes of repeated games, for *any* belief profile there are *no* strategy subsets that simultaneously satisfy learnability,

richness, and consistency. Thus, for example, if each player believes the other is playing a strategy that has a k-period memory, then one can show that learnability and richness hold but consistency fails: best responding in this setting requires using a strategy with a memory of more than k periods. The impossibility result generalizes to ε optimization and ε consistency, for ε sufficiently small. The result also generalizes to games with payoff uncertainty (with learnability, richness, and consistency now defined in terms of distributional strategies) (see Nachbar, 2001).

I conclude with four remarks. First, since the set of all strategies satisfies richness and consistency, it follows that the set of all strategies is not learnable for *any* beliefs: for any belief profile there is a strategy profile that the players will not learn to predict. This can also be shown directly by a diagonalization argument along the lines of Oakes (1985) and Dawid (1985). The impossibility result of Nachbar (2005) can be viewed as a game theoretic version of Dawid (1985). For a description of what subsets *are* learnable, see Noguchi (2005).

Second, if one constructs a Bayesian learning model satisfying learnability and consistency then LBR holds and, if players play their LBR strategies, play converges to equilibrium play. This identifies a potentially attractive class of Bayesian models in which convergence obtains. The impossibility result says, however, that if learnability and consistency hold, then player beliefs must be partially equilibrated in the sense of, in effect, excluding some of the strategies required by richness.

Third, consistency is not *necessary* for LBR or convergence. For example, for many stage games, variants of fictitious play satisfy LBR and converge even though these learning models are inconsistent. The impossibility result is a statement about the ability to construct Bayesian models with certain properties; it is not a statement about convergence per se.

Last, learnability, richness, and consistency may be too strong to be taken as necessary conditions for beliefs to be considered sensible. It is an open question whether one can construct Bayesian models satisfying conditions that are weaker but still strong enough to be interesting.

See Also

- ▶ Deterministic evolutionary dynamics
- ▶ Learning and evolution in games: adaptive heuristics
- ▶ Learning and evolution in games: an overview
- ▶ Learning and evolution in games: ESS
- ▶ Purification
- ▶ Repeated games
- ▶ Stochastic adaptive dynamics

Bibliography

Aoyagi, M. 1996. Evolution of beliefs and the Nash equilibrium of normal form games. *Journal of Economic Theory* 70: 444–469.

Bernheim, B.D. 1984. Rationalizable strategic behavior. *Econometrica* 52: 1007–1028.

Brown, G.W. 1951. Iterative solutions of games by fictitious play. In *Activity analysis of production and allocation*, ed. T.J. Koopmans. New York: Wiley.

Cournot, A. 1838. Researches into the Mathematical Principles of the Theory of Wealth. Trans. N.T. Bacon, New York: Kelley, 1960.

Dawid, A.P. 1985. The impossibility of inductive inference. *Journal of the American Statistical Association* 80: 340–341.

Foster, D., and P. Young. 2001. On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences* 98: 12848–12853.

Foster, D., and P. Young. 2003. Learning, hypothesis testing, and Nash equilibrium. *Games and Economic Behavior* 45: 73–96.

Fudenberg, D., and D. Kreps. 1993. Learning mixed equilibria. *Games and Economic Behavior* 5: 320–367.

Fudenberg, D., and D. Levine. 1993. Steady state learning and Nash equilibrium. *Econometrica* 61: 547–574.

Fudenberg, D., and D. Levine. 1995. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19: 1065–1089.

Fudenberg, D., and D. Levine. 1998. *Theory of learning in games*. Cambridge, MA: MIT Press.

Fudenberg, D., and D. Levine. 1999. Conditional universal consistency. *Games and Economic Behavior* 29: 104–130.

Harsanyi, J. 1973. Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory* 2: 1–23.

Hart, S., and A. Mas-Colell. 2003. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review* 93: 1830–1836.
Hart, S., and A. Mas-Colell. 2006. Stochastic uncoupled dynamics and Nash equilibrium. *Games and Economic Behavior* 57: 286–303.
Hofbauer, J., and W. Sandholm. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70: 2265–2294.
Jackson, M., and E. Kalai. 1999. False reputation in a society of players. *Journal of Economic Theory* 88: 40–59.
Jordan, J.S. 1991. Bayesian learning in normal form games. *Games and Economic Behavior* 3: 60–81.
Jordan, J.S. 1993. Three problems in learning mixed-strategy Nash equilibria. *Games and Economic Behavior* 5: 368–386.
Jordan, J.S. 1995. Bayesian learning in repeated games. *Games and Economic Behavior* 9: 8–20.
Kalai, E., and E. Lehrer. 1993a. Rational learning leads to Nash equilibrium. *Econometrica* 61: 1019–1045.
Kalai, E., and E. Lehrer. 1993b. Subjective equilibrium in repeated games. *Econometrica* 61: 1231–1240.
Kuhn, H.W. 1964. Extensive games and the problem of information. In *Contributions to the theory of games*, ed. M. Dresher, L.S. Shapley, and A.W. Tucker, Vol. 2. Princeton: Princeton University Press.
Milgrom, P., and J. Roberts. 1991. Adaptive and sophisticated learning in repeated normal form games. *Games and Economic Behavior* 3: 82–100.
Milgrom, P., and R. Weber. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10: 619–632.
Nachbar, J.H. 2001. Bayesian learning in repeated games of incomplete information. *Social Choice and Welfare* 18: 303–326.
Nachbar, J.H. 2005. Beliefs in repeated games. *Econometrica* 73: 459–480.
Noguchi, Y. 2005. Merging with a set of probability measures: A characterization. Working paper, Kanto Gakuin University.
Nyarko, Y. 1998. Bayesian learning and convergence to Nash equilibria without common priors. *Economic Theory* 11: 643–655.
Oakes, D. 1985. Self-calibrating priors do not exist. *Journal of the American Statistical Association* 80: 339–342.
Sandroni, A. 1998. Necessary and sufficient conditions for convergence to Nash equilibrium: the almost absolute continuity hypothesis. *Games and Economic Behavior* 22: 121–147.
Sandroni, A. 2000. Reciprocity and cooperation in repeated coordination games: The principled-player approach. *Games and Economic Behavior* 32: 157–182.
Shapley, L. 1962. On the nonconvergence of fictitious play. Discussion Paper RM-3026, RAND.

Learning and Evolution in Games: ESS

Ross Cressman

Abstract

The ESS concept, developed in the 1970s to predict through static fitness comparisons the evolutionary outcome of individual behaviours in a biological species, emerged as the cornerstone of evolutionary game theory. This theory is now as central to the analysis of strategic interactions in the social and management sciences as in the life sciences. The ESS also addresses stability questions for dynamics describing how individual behaviours evolve over time. Here, we summarize ESS theory as originally developed for symmetric two-player games and then discuss generalizations to population games, extensive form games, games with continuous strategy spaces, asymmetric and bimatrix games.

Keywords

Asymmetric games; Asymptotic stability; Backward induction; Best response dynamics; Bimatrix games; Buyer–seller game; Common interest games; Continuously stable strategy; Direct ESS; ESS (evolutionarily stable strategy); ESSet; Evolutionary dynamics; Evolutionary game theory; Extensive form games; Invasion barrier; Local superiority; Nash demand game; Nash equilibrium (NE); NE component; Neighbourhood invader strategy; Neighbourhood superiority; Neutrally stable strategies; Normal form games; Owner–intruder game; Partnership games; Pervasive strategy; Playing-the-field models; Population games; Prisoner's dilemma game; Replicator equation; Rock–scissors–paper game; Selten, R.; Strictly stable game; Symmetrical games; Two-species ESS; War of attrition game

JEL Classifications
C7

Introduction

According to John Maynard Smith in his influential book *Evolution and the Theory of Games* (1982, p.10), an ESS (that is, an *evolutionarily stable strategy*) is 'a strategy such that, if all members of the population adopt it, then no mutant strategy could invade the population under the influence of natural selection'. The ESS concept, based on static fitness comparisons, was originally introduced and developed in the biological literature (Maynard Smith and Price 1973) as a means to predict the eventual outcome of evolution for individual behaviours in a single species. It avoids the complicated dynamics of the evolving population that may ultimately depend on spatial, genetic and population size effects.

To illustrate the Maynard Smith (1982) approach, suppose individual fitness is the expected payoff in a random pairwise contest. The ESS strategy p^* must then do at least as well as a mutant strategy p in their most common contests against p^* and, if these contests yield the same payoff, then p^* must do better than p in their rare contests against a mutant. That is, Maynard Smith's definition applied to a symmetric two-player game says p^* is an ESS if and only if, for all $p \neq p^*$,

$$\begin{aligned} &(i) \quad \pi(p,p^*) \leq \pi(p^*,p^*) \ \text{(equilibriumcondition)} \\ &(ii) \quad \text{if} \pi(p,p^*) = \pi(p^*,p^*), \\ &\quad \pi(p,p) < \pi(p^*,p) \quad \text{(stabilitycondition)} \end{aligned} \tag{1}$$

where $\pi(p,\hat{p})$ is the payoff of p against $\hat{p}$. One reason the ESS concept has proven so durable is that it has equivalent formulations that are equally intuitive (see especially the concepts of invasion barrier and local superiority in Section "Normal Form Games").

By (1) (i), an ESS is a Nash equilibrium (NE) with the extra refinement condition (ii) that seems heuristically related to dynamic stability. In fact, there is a complex relationship between the static ESS conditions and dynamic stability, as illustrated throughout this article with specific reference to the replicator equation. It is this relationship that formed the initial basis of what has come to be known as 'evolutionary game theory'.

ESS theory (and evolutionary game theory in general) has been extended to many classes of games besides those based on a symmetric two-player game. This article begins with ESS theory for symmetric normal form games before briefly describing the additional features that arise in each of several types of more general games. The unifying principle of local (or neighborhood) superiority will emerge in the process.

ESS for Symmetric Games

In a symmetric evolutionary game, there is a single set S of pure strategies available to the players, and the payoff to pure strategy e_i is a function π_i of the system's strategy distribution. In the following subsections we consider two-player symmetric games with S finite in normal and extensive forms (Sections "Normal Form Games" and "Extensive Form Games" respectively) and with S a continuous set (Section "Continuous Strategy Space").

Normal Form Games

Let $S \equiv \{e_1, \ \ldots, e_n\}$ be the set of pure strategies. A player may also use a mixed strategy $p \in \Delta^n \equiv \{p = (p_1, \ \ldots, p_n) | \sum p_i = 1, p_i \geq 0\}$ where p_i is the proportion of the time this individual uses pure strategy e_i. Pure strategy e_i is identified with the ith unit vector in Δ^n. The population state is $\hat{p} \in \Delta^n$ whose components are the current frequencies of strategy use in the population (that is, the strategy distribution). We assume the expected payoff to p is the bilinear function $\pi(p,\hat{p}) = \sum_{i,j=1}^{n} p_i \pi(e_i,e_j)\hat{p}_j$ resulting from random two-player contests.

Suppose the resident population is monomorphic at p^* (that is, all members adopt strategy p^*) and a monomorphic sub-population of mutants using p appears in the system. These mutants

will not invade if there is a positive *invasion barrier* $\varepsilon_0(p)$ (Bomze and Pötscher 1989). That is, if the proportion ε of mutants in the system is less than $\varepsilon_0(p)$, then the mutants will eventually die out due to their lower replication rate. In mathematical terms, $\varepsilon = 0$ is a (locally) asymptotically stable rest point of the corresponding resident-mutant invasion dynamics. For invasion dynamics based on replication, Bomze and Pötscher show p^* is an ESS (that is, satisfies (1)) if and only if every $p \neq p^*$ has a positive invasion barrier.

Important and somewhat surprising consequences of an ESS p^* are its asymptotic stability for many evolutionary dynamics beyond these monomorphic resident systems invaded by a single type of mutant. For instance, p^* is asymptotically stable when simultaneously invaded by several types of mutants and when a polymorphic resident system consisting of several (mixed) strategy types whose average strategy is p^* is invaded (see the 'strong stability' concept developed in Cressman 1992). In particular, p^* is asymptotically stable for the replicator equation (Taylor and Jonker 1978; Hofbauer et al. 1979; Zeeman 1980)

$$\dot{p}_i = p_i(\pi(e_i, p) - \pi(p, p)) \tag{2}$$

when each individual player is a pure strategist.

Games that have a completely mixed ESS (that is, p^* is in the interior of Δ^n) enjoy further dynamic stability properties since these games are *strictly stable* (that is, $\pi(p - \hat{p}, p - \hat{p}) < 0$ for all $p \neq \hat{p}$) (Sandholm 2006). The ESS of a strictly stable game is also globally asymptotically stable for the best response dynamics (the continuous-time version of fictitious play) (Hofbauer and Sigmund 1998) and for the Brown–von Neumann–Nash dynamics (related to Nash's 1951, proof of existence of NE) (Hofbauer and Sigmund 2003).

The preceding two paragraphs provide a strong argument that an ESS will be the ultimate outcome of the evolutionary adjustment process. The proofs of these results use two other equivalent characterizations of an ESS p^* of a symmetric normal form game; namely,

(a) p^* has a *uniform* invasion barrier (i.e. $\varepsilon_0(p) > 0$ is independent of p)
(b) for all p sufficiently close (but not equal) to p^*

$$\pi(p, p) < \pi(p^*, p). \tag{3}$$

It is this last characterization, called 'local superiority' (Weibull 1995), that proves so useful for other classes of games (see below). Heuristically, (3) suggests p^* will be asymptotically stable since there is an incentive to shift towards p^* whenever the system is slightly perturbed from p^*. Unfortunately, there are many normal form games that have no ESS. These include most three-strategy games classified by Zeeman (1980) and Bomze (1995). No mixed strategy p^* can be an ESS of a symmetric zero-sum game (that is, $\pi(\hat{p}, p) = -\pi(p, \hat{p})$ for all p, $\hat{p} \in \Delta^n$) since $\pi(p^*, p) = \pi(-p^* - p, p) \leq 0 = \pi(p, p)$ for all $p \in \Delta^n$ in some direction from p^*. Thus, the classic zero-sum Rock–Scissors–Paper Game in Table 1 has no ESS since its only NE $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is interior. An early attempt to relax the ESS conditions to rectify this replaces the strict inequality in (1) (ii) by $\pi(p, p) \leq \pi(p^*, p)$. The NE p^* is then called a *neutrally stable strategy* (NSS) (Maynard Smith 1982, Weibull 1995). The only NE of the Rock–Scissors–Paper Game is a NSS.

The Payoff Matrix for the Rock–Scissors–Paper Game

$$\begin{array}{c} \\ \text{Rock} \\ \text{Scissors} \\ \text{Paper} \end{array} \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

Each entry is the payoff to the row player when column players are listed in the same order.

Also, the normal forms of most interesting extensive form games have no ESS, especially when NE outcomes do not specify choices off the equilibrium path and so correspond to NE components. In general, when NE are not isolated,

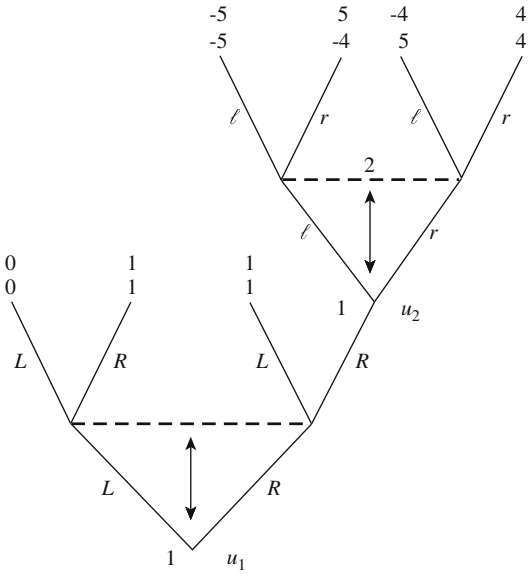
the *ESSet* introduced by Thomas (1985) is more important. This is a set E of NSS so that (1) (ii) holds for all $p^* \in E$ and $p \notin E$. An ESSet is a finite union of disjoint NE components, each of which must be an ESSet in its own right. Each ESSet has setwise dynamic stability consequences analogous to an ESS (Cressman 2003). The ES structure of a game refers to its collection of ESSs and ESSets.

There are then several classes of symmetric games that always have an ESSet. Every two-strategy game has an ESSet (Cressman 2003) which generically (that is, unless $\pi(\hat{p},\hat{p}) = \pi(p,\hat{p})$ for all $p,\hat{p} \in \Delta^2$ is a finite set of ESSs. All games with symmetric payoff function (that is, $\pi(\hat{p},p) = \pi(p,\hat{p})$ for all p, $\hat{p} \in \Delta^n$) have an ESSet corresponding to the set of local maxima of $\pi(p, p)$ which generically is a set of isolated ESSs). These are called partnership games (Hofbauer and Sigmund 1998) or common interest games (Sandholm 2006).

Symmetric games with payoff, $\pi_i(\hat{p})$, of pure strategy e_i nonlinear in the population state $\hat{p}$ are quite common in biology and in economics (Maynard Smith 1982; Sandholm 2006), where they are called playing-the-field models or population games. With $\pi_i(p,\hat{p}) = \sum_i p_i \pi_i(\hat{p})$, nonlinearity implies (1) is a weaker condition than (3), as examples in Bomze and Pötscher (1989) show. Local superiority (3) is then taken as the operative definition of an ESS p^* (Hofbauer and Sigmund 1998) and it is equivalent to the existence of a uniform invasion barrier for p^*.

Extensive Form Games

The application of ESS theory to finite extensive form games has been less successful (see Fig. 1). Every ESS can have no other realization equivalent strategies in its normal form (van Damme 1991) and so, in particular, must be *pervasive strategy* (that is, it must reach every information set when played against itself). To ease these problems, Selten (1983) defined a *direct ESS* in terms of behaviour strategies (that is, strategies that specify the local behaviour at each player information set) as a b^* that satisfies (1) for any other behaviour strategy b. He showed each such b^* is subgame perfect and arises from the backward induction technique applied to the ES structure of the subgames and their corresponding truncations.



Learning and Evolution in Games: ESS, Fig. 1 The extensive form tree of the van Damme example. For the construction of the tree of a symmetric extensive form game, see Selten (1983) or van Damme (1991)

Consider backward induction applied to Fig. 1. Its second-stage subgame $\begin{matrix}\ell \\ r\end{matrix}\begin{bmatrix}-5 & 5 \\ -4 & 4\end{bmatrix}$ has mixed ESS $b_2^* = \left(\frac{1}{2}, \frac{1}{2}\right)$ and, when the second decision point of player 1 is replaced by the payoff 0 from b^*, the truncated single-stage game $\begin{matrix}L \\ R\end{matrix}\begin{bmatrix}0 & 1 \\ 1 & 0\end{bmatrix}$ also has a mixed ESS $b_1^* = \left(\frac{1}{2}, \frac{1}{2}\right)$. Since both stage games have a mixed ESS (and so a unique NE since they are strictly stable), $\left(b_1^*, b_2^*\right)$ is the only NE of Fig. 1 and it is pervasive. Surprisingly, this example has no direct ESS as Selten originally hoped since $\left(b_1^*, b_2^*\right)$ can be invaded by the pure strategy that plays *Rr* (van Damme 1991).

The same technique applied to Fig. 1 with second-stage subgame replaced by

$\begin{matrix}\ell \\ r\end{matrix}\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}$ yields $b_2^* = \left(\frac{1}{2}, \frac{1}{2}\right)$ and truncated single-stage game $\begin{matrix}L \\ R\end{matrix}\begin{bmatrix}0 & 1\\ 1 & -1/2\end{bmatrix}$ with $b_1^* = \left(\frac{3}{5}, \frac{2}{5}\right)$. This is an example of a two-stage War of Attrition with base game $\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$ where a player remains (R) at the first stage in the hope the opponent will leave (L) but incurs a waiting cost of one payoff unit if both players remain. This $\left(b_1^*, b_2^*\right)$ is a direct ESS since all N-stage War of Attrition games are strictly stable (Cressman 2003).

The examples in the preceding two paragraphs show that, although backward induction determines candidates for the ES structure, it is not useful for determining which candidates are actually direct ESSs. The situation is more discouraging for non-pervasive NE. For example, the only NE outcome of the two-stage repeated Prisoner's Dilemma game (Nachbar 1992) with cumulative payoffs is mutual defection at each stage. This NE outcome cannot be an isolated behaviour strategy (that is, there is a corresponding NE component) and so there is no direct ESS. Worse, for typical single-stage payoffs such as, $\begin{matrix}\text{Defect} \\ \text{Cooperate}\end{matrix}\begin{bmatrix}-1 & 10\\ -2 & 5\end{bmatrix}$ this component does not satisfy setwise extensions of the ESS (for example, it is not an ESSet).

Characterization of NE found by backward induction with respect to dynamically stable rest points of the subgames and their truncations shows more promise. Each direct ESS b^* yields an ESSet in the game's normal form (Cressman 2003) and so is dynamically stable. Furthermore, for the class of simultaneity games where both players know all player actions at earlier stages, Cressman shows that, if b^* is a pervasive NE, then it is asymptotically stable with respect to the replicator equation if and only if it comes from this backward induction process. In particular, the NE for Fig. 1 and for the N-stage War of Attrition are (globally) asymptotically stable. Although the subgame perfect NE for the N-stage Prisoner's Dilemma game that defects at each decision point is not asymptotically stable, the eventual outcome of evolution is in the NE component (Nachbar 1992; Cressman 2003).

Continuous Strategy Space

Evolutionary game theory for symmetric games with a continuous set of pure strategies S has been slower to develop. Most recent work examines static payoff comparisons that predict an $x^* \in S$ is the evolutionary outcome. There are now fundamental differences between the ESS notion (1) and that of local superiority (3) as well as between invasion by monomorphic mutant sub-populations and the polymorphic model of the replicator equation. Here, we illustrate these differences when S is a subinterval of real numbers and $\pi(x, y)$ is a continuous payoff function of $x, y \in S$.

First, consider an $x^* \in S$ that satisfies (3). In particular,

$$\pi(x, x) < \pi(x^*, x) \tag{4}$$

for all $x \in S$ sufficiently close (but not equal) to x^*. This is the *neighbourhood invader strategy* (NIS) condition of Apaloo (1997) that states x^* can invade any nearby monomorphism x. On the other hand, from (1), x^* cannot be invaded by these x if it is a *neighbourhood strict NE*, that is

$$\pi(x, x^*) < \pi(x^*, x^*) \tag{5}$$

for any other x sufficiently close to x^*. Inequalities (4) (5) are independent of each other and combine to assert that x^* strictly dominates x in all these two-strategy games $\{x^*, x\}$.

In the polymorphic model, populations are described by a P in the infinite dimensional set $\Delta(S)$ of probability distributions with support in S. When the expected payoff $\pi(x, P)$ is given through random pairwise contests, Cressman (2005) shows that strict domination implies x^* is *neighbourhood superior* (that is,

$$\pi(x^*, P) > \pi(P, P) \tag{6}$$

for all other $P \in \Delta(S)$ with support sufficiently close to x^*) and conversely, neighbourhood

superiority implies weak domination. Furthermore, a neighborhood superior monomorphic population x^* (that is, the Dirac delta probability distribution δx^*) is asymptotically stable for all initial P with support sufficiently close to x (and containing x^*) under the replicator equation. This is now a dynamic on $\Delta(S)$ (Oechssler and Riedel 2002) that models the evolution of the population distribution.

In the monomorphic model, the population is a monomorphism $x(t) \in S$ at all times. If a nearby mutant strategy $y \in S$ can invade x, the whole population is shifted in this direction. This intuition led Eshel (1983) to define a *continuously stable strategy* (CSS) as a neighbourhood strict NE x^* that satisfies, for all x sufficiently close to x^*,

$$\pi(y,x) > \pi(x,x) \tag{7}$$

for all y between x^* and x that are sufficiently close to x. Later, Dieckmann and Law (1996) developed the canonical equation of adaptive dynamics to model the evolution of this monomorphism and showed a neighbourhood strict NE x^* is a CSS if and only if it is an asymptotically stable rest point. Cressman (2005) shows x^* is a CSS if and only if it is *neighbourhood half-superior* (that is, there is a uniform invasion barrier of at least $\frac{1}{2}$ in the two-strategy games $\{x^*, x\}$) (see also the half-dominant concept of Morris et al. 1995).

For example, take $S = \mathbf{R}$ and payoff function

$$\pi(x,y) = -x^2 + bxy \tag{8}$$

that has strict NE $x^* = 0$ for all values of the fixed parameter b. x^* is a NIS (CSS) if and only if $b < 1$ ($b < 2$) (Cressman and Hofbauer, 2005). Thus, there are strict NE when $b > 2$ that are not 'evolutionarily stable'.

Asymmetric Games

Following Selten (1980) and van Damme (1991), in a two-player asymmetric game with two roles (or species), pairwise contests may involve players in the same or in opposite roles. First, consider ESS theory when there is a finite set of pure strategies $S = \{e_1, \ldots, e_n\}$ and $T = \{f_1, \ldots, f_m\}$ for players in role 1 and 2 respectively. Assume payoff to a mixed strategist is given by a bilinear payoff function and let $\pi_1(p,\hat{p},\hat{q})$ be the payoff to a player in role one using $p \in \Delta^n$ when the current state of the population in roles 1 and 2 are $\hat{p}$ and $\hat{q}$ respectively. Similarly, $\pi_2(q,\hat{p},\hat{q})$ is the payoff to a player in role 2 using $q \in \Delta^m$. For a discussion of resident-mutant invasion dynamics, see Cressman (1992), who shows the monomorphism (p^*, q^*) is uninvadable by any other mutant pair (p, q) if and only if it is a *two-species ESS*, that is, for all (p, q) sufficiently close (but not equal) to (p^*, q^*),

$$\text{either } \pi_1(p;p,q) < \pi_1(p^*;p,q) \text{ or } \pi_2(q;p,q) < \pi_2(q^*;p,q). \tag{9}$$

The ESS condition (9) is the two-role version of local superiority (3) and has an equivalent formulation analogous to (1) (Cressman 1992). This ESS also enjoys similar stability properties to the ESS of Subsection "Normal Form Games" such as its asymptotic stability under the (two-species) replicator equation (Cressman 1992, 2003).

A particularly important class of asymmetric games consists of truly asymmetric games that have no contests between players in the same role (that is, there are no intraspecific contests). These are bimatrix games (that is, given by an $n \times m$ matrix whose ijth entry is the pair of payoffs $(\pi_1(e_i, f_j), \pi_2(e_i, f_j))$ for the interspecific contest between e_i and f_j). The ESS concept is now quite restrictive since Selten (1980) showed that (p^*, q^*) satisfies (9) if and only if it is a strict NE. This is also equivalent to asymptotic stability under the (two-species) replicator equation (Cressman 2003). Standard examples (Cressman 2003), with two strategies for each player include the Buyer–Seller Game that has no ESS since its only NE is in the interior.

Another is the Owner–Intruder Game that has two strict NE Maynard Smith (1982) called the bourgeois ESS where the owners defend their territory and the paradoxical ESS where owners retreat.

Asymmetric games with continuous sets of strategies have recently received a great deal of attention (Leimar 2006). For a discussion of neighbourhood (half) superiority conditions that generalize (6) and (7) to two-role truly asymmetric games with continuous payoff functions, see Cressman (2005). He also shows how these conditions are related to NIS and CSS concepts based on (9) and to equilibrium selection results for games with discontinuous payoff functions such as the Nash Demand Game (Binmore et al. 2003).

See Also

- ▶ Deterministic Evolutionary Dynamics
- ▶ Learning and Evolution in Games: An Overview

Bibliography

Apaloo, J. 1997. Revisiting strategic models of evolution: The concept of neighborhood invader strategies. *Theoretical Population Biology* 52: 71–77.

Binmore, K., L. Samuelson, and P. Young. 2003. Equilibrium selection in bargaining models. *Games and Economic Behavior* 45: 296–328.

Bomze, I. 1995. Lotka–Volterra equation and replicator dynamics: New issues in classification. *Biological Cybernetics* 72: 447–453.

Bomze, I., and B. Pötscher. 1989. *Game theoretical foundations of evolutionary stability*. Lecture notes in economics and mathematical systems 324. Berlin: Springer-Verlag.

Cressman, R. 1992. *The stability concept of evolutionary games (A dynamic approach)*. Lecture notes in biomathematics 94. Berlin: Springer-Verlag.

Cressman, R. 2003. *Evolutionary dynamics and extensive form games*. Cambridge, MA: MIT Press.

Cressman, R. 2005. Continuously stable strategies, neighborhood superiority and two-player games with continuous strategy space. Mimeo.

Cressman, R., and J. Hofbauer. 2005. Measure dynamics on a one-dimensional continuous trait space: Theoretical foundations for adaptive dynamics. *Theoretical Population Biology* 67: 47–59.

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

Eshel, I. 1983. Evolutionary and continuous stability. *Journal of Theoretical Biology* 103: 99–111.

Hofbauer, J., P. Schuster, and K. Sigmund. 1979. A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 81: 609–612.

Hofbauer, J., and K. Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.

Hofbauer, J., and K. Sigmund. 2003. Evolutionary game dynamics. *Bulletin of the American Mathematical Society* 40: 479–519.

Leimar, O. 2006. Multidimensional convergence stability and the canonical adaptive dynamics. In *Elements of adaptive dynamics*, ed. U. Dieckmann and J. Metz. Cambridge: University Press.

Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.

Maynard Smith, J., and G. Price. 1973. The logic of animal conflicts. *Nature* 246: 15–18.

Morris, S., R. Rob, and H. Shin. 1995. Dominance and belief potential. *Econometrica* 63: 145–157.

Nachbar, J. 1992. Evolution in the finitely repeated Prisoner's Dilemma. *Journal of Economic Behavior and Organization* 19: 307–326.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.

Oechssler, J., and F. Riedel. 2002. On the dynamic foundation of evolutionary stability in continuous models. *Journal of Economic Theory* 107: 223–252.

Sandholm, W. 2006. *Population games and evolutionary dynamics*. Cambridge, MA: MIT Press.

Selten, R. 1980. A note on evolutionarily stable strategies in asymmetrical animal contests. *Journal of Theoretical Biology* 84: 93–101.

Selten, R. 1983. Evolutionary stability in extensive two-person games. *Mathematical Social Sciences* 5: 269–363.

Taylor, P., and L. Jonker. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40: 145–156.

Thomas, B. 1985. On evolutionarily stable sets. *Journal of Mathematical Biology* 22: 105–115.

van Damme, E. 1991. *Stability and perfection of nash equilibria*. 2nd ed. Berlin: Springer-Verlag.

Weibull, J. 1995. *Evolutionary game theory*. Cambridge, MA: MIT Press.

Zeeman, E. 1980. Population dynamics from game theory. In *Global theory of dynamical systems*, ed. Z. Nitecki and C. Robinson. Lecture notes in mathematics 819. Berlin: Springer.

Learning and Information Aggregation in Networks

Douglas Gale and Shachar Kariv

Abstract

'Social learning' is a process whereby economic agents learn by observing the behaviour of others. 'Social learning in networks' requires sophistication because individuals draw inferences from the behaviour of agents they cannot directly observe. Theoretical research suggests that, even if networks are very incomplete, social learning leads to uniform behaviour. Experimental evidence suggests that learning in networks conforms quite well to theoretical predictions. It also illustrates how the network architecture influences the pattern of learning and the efficiency of information aggregation.

Keywords

Bala–Goyal model; Bounded rationality; Circle network; Complete network; Connected graph; Directed graph; Griliches, Z.; Herd behavior; Hubs; Imitation principle; Information aggregation; Informational cascades; Perfect information; Pure information externality; Scale-free networks; Social experimentation; Social learning; Star network

JEL Classifications

D85

'Social learning' is a process whereby economic agents learn by observing the actions (but not the payoffs) of others; 'social learning in networks' applies this idea to situations in which individuals observe the other individuals to whom they are connected in a social network.

Griliches (1957) first studied the gradual adoption of corn planted with hybrid seed in the USA, a new agricultural technique, from the early 1930s to mid-1950s. He observed that at first farmers learned from salespersons; later they learned from their neighbours. The result was an S-shaped time profile of adoption. A number of recent papers, including Foster and Rosenzweig (1995), Conley and Udry (2001), Kremer and Miguel (2003) and Munshi (2004) examine how agents in developing countries learn from their social contacts when deciding whether to adopt new technologies.

The classical model of social learning, first studied by Banerjee (1992) and Bikhchandani et al. (1992), and extended by Smith and Sørensen (2000), assumes a *pure information externality.* An agent's payoff $u(a,w)$ depends only on his own action a and an unknown state of nature w. Each agent i has private information about the state and his choice of action a will reflect that information. By observing an agent's action, it is possible to learn something about his information and make a better decision. The problem is that agents may rationally ignore their own information and 'follow the herd', that is, imitate the actions they see others choose. So-called *herd behaviour* and *informational cascades* can arise very rapidly, before much information has been revealed, and often result in inefficient choices. A number of experimental studies replicate herd behaviour in the laboratory.

The classical models assume that agents make decisions sequentially and observe the action chosen by each of their predecessors. In reality, individuals are bound together by a *social network,* the complex of relationships that brings them into contact with other agents, such as neighbours, co-workers, family, and so on. A specific framework, introduced by Gale and Kariv (2003), henceforth GK, assumes that individuals are bound together by a social network and can observe the agents to whom they are connected only through the network. The social network is represented by a *directed graph* in which nodes correspond to agents and agent i can observe agent j if there is an edge leading from node i to node j. In order to model the diffusion of

information through the network, GK assume that agents choose actions simultaneously and revise their decisions as new information is received. More precisely, an agent whose current information is I chooses an action a to maximize his short-run payoff $E[u(a,w)|I]$. GK rationalize non-strategic behaviour by assuming there is a large number of agents of each type, so a single agent's decision has no impact on the future play of the game.

An agent's beliefs can be represented by a random sequence of probability distributions $P_t(\omega)$. At date t, an agent derives a posterior $P_{t+1}(w)$ from the prior $P_t(w)$ and the new information received. These beliefs satisfy the martingale property $E[P_{t+1}|I_t] = P_t$, and the martingale convergence theorem implies that these beliefs converge to a constant with probability one. The limiting beliefs are not necessarily uniform (different agents may have different beliefs) and need not be fully revealing. However, in *connected* networks, where every agent is connected directly or indirectly with every other agent, the initial diversity of actions is eventually replaced by uniformity. More precisely, except in cases of indifference, agents will choose the same action. This is the network-learning analogue of the herd behaviour found in the classical social learning model. The proof of uniformity makes use of the *imitation principle.* If agent i can observe the actions of agent j, agent i must be able to do as well as j on average (because one feasible strategy is to choose the same action j). In a connected network, all agents get the same payoff on average and this implies that they choose different actions only if they are indifferent.

Learning in a network is 'simply' a matter of Bayesian updating, but a rational agent must take account of the network architecture in order to update correctly. For example, suppose there are three (types of) agents, *A, B,* and *C*, arranged in a circle: *A* observes *B*, *B* observes *C*, and *C* observes *A*. At the first decision, *A* has not yet had a chance to observe *B*, so he makes his decision based on his private information. Before the second decision, *A* observes *B*'s first decision and uses it to update his beliefs about the true state of nature. Before the third decision, *A* observes *B*'s second decision and realizes that any change from the first must be based on *B*'s observation of *C*'s first decision. So now *A* can make some inference about *C*'s private information and update his beliefs accordingly. This learning can go on for some time. Eventually, *A* may observe changes in *B*'s action that were prompted by changes in *C*'s action that were prompted by *C*'s observation of *A*. Even this is informative because it reveals how strong *C*'s information is relative to *A*'s. In any case, exploiting fully the information revealed in a network requires agents to consider not only what they observe, but also what their neighbours observe, what their neighbours' neighbors observe, and so on. The chains of inferences that rational individuals make naturally involve hierarchies of beliefs, that is, beliefs about a neighbour's beliefs about a neighbour's beliefs about. . ..

The complexity of Bayesian learning in networks has led some authors to suggest that models of *bounded rationality* are more appropriate for describing learning in networks. Bala and Goyal (1998) examine the decisions of boundedly rational agents, who try to extract information from the actions and payoffs of the agents they observe, but without taking account of the fact that those agents also observe other agents. Hence, there is private information in the Bala–Goyal model, but agents are assumed to ignore it. In the Bala–Goyal model, at each date, an agent chooses one of several available actions with unknown payoff distributions. Agents can observe the actions and payoffs of their neighbours (those to whom they are directly connected by the network) and use this information to update their beliefs about the payoff distribution. Thus, agents learn by observing the outcome (payoff) of an experiment (choice of action) rather than by inferring another agent's private information from his action. This is a model of *social experimentation*, in the sense that it generalizes the problem of a single agent experimenting with a multi-armed bandit to a social setting, rather than social learning. A model of social experimentation is quite different from a model of social learning because there is an

informational externality but there is no informational asymmetry. As with Bayesian learning, boundedly rational learning implies convergence of beliefs and uniformity of actions in the limit.

Laboratory experiments provide the cleanest test for the theory since subjects' neighbourhoods and private information can be controlled. Choi et al. (2004, 2005) describe the results of an experimental investigation of learning in networks based on the model of GK. The experiments involve three-person, connected social networks. The experimental design uses three representative networks: the *complete network*, in which each agent can observe the actions chosen by the other agents; the *star network*, in which one agent, the centre, can observe the actions of the other two peripheral agents, and the peripheral agents can observe only the centre; and the *circle network*, in which each agent can observe only one other agent and each agent is observed by one other agent. Despite the small number of players in each game, it can be shown that myopic payoff maximization is rational: there is no gain to strategic behaviour. Nonetheless, larger-scale experiments might be informative.

The experimental data from these studies exhibit a strong tendency toward herd behaviour, but despite this tendency the efficiency of information aggregation is quite good. Although convergence to a uniform action is quite rapid, frequently occurring within two to three turns, there are significant differences between the behaviour of different networks. Most herds entail correct decisions, which is consistent with the predictions of the parametric model underlying the experimental design. Comparing the behaviour of different individuals indicates that there is indeed high variation in individual behaviour across subjects, but the error rates (the proportion of times a subject deviates from the best response) are uniformly fairly low.

These results suggest that the theory adequately accounts for large-scale features of the data, but in some situations the theory does less well in accounting for subjects' behaviour. It is likely that the theory fails in those situations because the complexity of the decision problem exceeds the bounded rationality of the subjects. Clearly, because of the lack of common knowledge in the networks, the decision problems faced by subjects require quite sophisticated reasoning. Subjects' success or failure in the experiment results from the appropriateness of the heuristics they use as much as the inherent difficulty of the decision-making. Thus, an important subject for future research is to identify 'black spots' where the theory does least well in interpreting the data and ask whether additional 'behavioural' explanations might be needed to account for the subjects' behaviour.

Many important questions about social learning in networks remain to be explored. While small networks can be very insightful, especially in experimental contexts, the development of the theory depends on properties of networks that can be generalized. The recent discovery of Barabási and Albert (1999) that many networks are *scale-free,* in the sense that a few nodes are *hubs,* which have a very large number of links to other nodes whereas most nodes have just a few, has significant implications for the efficiency of information aggregation. Once information reaches a hub it passes to numerous other nodes and spreads rapidly throughout the entire population, but if the hub's information is of poor quality its disproportionate influence becomes a disadvantage. Thus, the impact of hubs on the efficiency of information aggregation is not clear. Perhaps the most important subject for future research is to identify the impact of network architecture on the efficiency and dynamics of social learning. Progress in this area requires both new theory and new experimental data.

See Also

- ▶ Behavioural Game Theory
- ▶ Experimental Economics, History of
- ▶ Griliches, Zvi (1930–1999)
- ▶ Information Cascades
- ▶ Learning and Evolution in Games: An Overview
- ▶ Logit Models of Individual Choice
- ▶ Network Formation

Bibliography

Bala, V., and S. Goyal. 1998. Learning from neighbors. *Review of Economic Studies* 65: 595–621.

Banerjee, A. 1992. A simple model of herd behavior. *Quarterly Journal of Economics* 107: 797–817.

Barabási, A., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286: 509–512.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascade. *Journal of Political Economy* 100: 992–1026.

Choi, S., D. Gale, and S. Kariv. 2004. *Learning in networks: An experimental study.* Mimeo. New York: Center for Experimental Social Science (CESS), New York University.

Choi, S., D. Gale, and S. Kariv. 2005. Behavioral aspects of learning in social networks: An experimental study. In *Experimental and behavioral economics advances in applied microeconomics*, vol. 13, ed. J. Morgan. Oxford: JAI Press.

Conley, T., and C. Udry. 2001. Social learning through networks: The adoption of new agricultural technologies in Ghana. *American Journal of Agricultural Economics* 83: 668–673.

Foster, A., and M. Rosenzweig. 1995. Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy* 103: 1176–1209.

Gale, D., and S. Kariv. 2003. Bayesian learning in social networks. *Games and Economic Behavior* 45: 329–346.

Griliches, Z. 1957. Hybrid corn: An exploration in the economics of technological change. *Econometrica* 25: 501–522.

Kremer, M., and T. Miguel. 2003. *Networks, social learning, and technology adoption: The case of deworming drugs in Kenya*. Mimeo. Berkeley: Department of Economics, University of California.

Munshi, K. 2004. Social learning in a heterogeneous population: Technology diffusion in the Indian Green Revolution. *Journal of Development Economics* 73: 185–213.

Smith, L., and P. Sørensen. 2000. Pathological outcomes of observational learning. *Econometrica* 68: 371–398.

Learning in Macroeconomics

George W. Evans and Seppo Honkapohja

Abstract

Expectations play a key role in macroeconomics. The assumption of rational expectations has been recently relaxed by explicit models of forecasting and model updating. Rational expectations can be assessed for stability under various types of learning, with least squares learning playing a prominent role. In addition to assessing the plausibility of an equilibrium, learning also provides a selection criterion when there are multiple equilibria. Monetary policy should be designed to avoid instability under learning and to facilitate coordination on desirable equilibria. Learning can also help to explain macroeconomic fluctuations as arising through either instabilities, stable indeterminacies or persistent learning dynamics.

Keywords

Actual law of motion; Adaptive learning; Animal spirits; Asset pricing; Bounded rationality; Cagan model; Calibration; Constant gain learning; Eductive learning; Expectational stability; Expectations; Game theory; Hyperinflation; Imperfect knowledge; Increasing social returns; Indeterminacy of equilibrium; Interest-rate rule; Learning in macroeconomics; Liquidity trap; Monetary policy; Multiple equilibria; New Keynesian macroeconomics; Overlapping contracts model; Overlapping generations model; Perceived law of motion; Phillips curve; Rational expectations; Rational expectations equilibrium; Rational learning; Real business cycles; Recursive least squares; Seigniorage; Stationary sunspot equilibria; Sunspot equilibrium; Taylor rule; Uniqueness of equilibrium; Vector autoregression

JEL Classifications

D4; D10

Learning in macroeconomics refers to models of expectation formation in which agents revise their forecast rules over time, for example in response to new data. Expectations of future income, prices and sales play key roles in theories of saving and investment. Many other examples of the central role of expectations could be given.

Introduction

The current standard methodology for modelling expectations is to assume that the economy is in a rational expectations equilibrium (REE). REE is a model-consistent equilibrium in the two-way relationship between the influence of expectations on the economy and the dependence of expectations on the time path of the economy.

The standard formulation of REE makes strong assumptions on the information of economic agents. The true stochastic process of the economy is assumed known, with unforecastable random shocks constituting the remaining uncertainty. This assumption presupposes that the economic agents know much more than, say, the economists who in practice do not know the true stochastic structure and instead must estimate its parameters.

Recently, macroeconomic theory has been moving beyond the strict rational expectations (RE) hypothesis. Explicit models of imperfect knowledge and associated learning processes have been developed. In models of learning economic agents try to improve their knowledge of the stochastic process of the economy over time as new information becomes available.

Different approaches to modelling learning behaviour have been employed. Perhaps the most common has been 'adaptive learning', which views economic agents as econometricians who estimate the parameters of their model and make forecasts using their estimates. In adaptive learning economic agents have limited common knowledge since they estimate their own perceived laws of motion.

A second approach, called 'eductive learning', assumes common knowledge of rationality: economic agents engage in a process of reasoning about the possible outcomes knowing that other agents engage in the same process. Eductive learning takes place in logical time. A third approach has been 'rational learning', which employs a Bayesian viewpoint. Full knowledge of economic parameters is then replaced by priors and Bayesian updating under a correctly specified model, including common knowledge that all agents share this knowledge. Rational learning thus retains a form of REE at each point of time.

Basic theories of learning were developed largely in the 1980s and 1990s. See Sargent (1993, 1999), Evans and Honkapohja (2001), Guesnerie (2005) and Beck and Wieland (2002) for references. Recently, models of learning have been applied to issues of macroeconomic, and especially monetary, policy. In this overview, we focus on adaptive learning as it has been the most widely used approach. (For references to the pre-2001 literature, see Evans and Honkapohja 2001.)

Least Squares Learning

In adaptive learning it is commonly assumed that agents estimate their model of the dynamics of economic variables, called the *perceived law of motion* (PLM), by *recursive least squares* (RLS), arguably the most common estimation method in econometrics.

Overview

We illustrate the key concepts using the Cagan model of the price level $\hat{m} - p_t = -\psi\left(p^e_{t+1} - p_t\right) + \phi' w_t + \varepsilon_t$, where p_t. and $\hat{m}$ are logarithms of the price level and (constant) nominal money supply. Here $\psi > 0$ and p^e_{t+1} enotes the expectations of p_{t+1} formed at time t. w_t is a vector of observable exogenous variables, assumed to follow a stationary vector autoregression (VAR) process $w_t = Fw_{t-1} + e_t$, in which F is taken as known for simplicity. ε_t is an unobservable i.i.d. shock.

The reduced form of the Cagan model is

$$p_t = \alpha_0 + \alpha_1 p^e_{t+1} + \beta' w_t + v_t, \qquad (1)$$

where $v_t = -(1+\psi)^{-1}\varepsilon_t$ and α_0, α_1 and β depend on $\hat{m}, \psi$ and ϕ. The model has a unique REE of the form $p_t = \bar{a} + \overline{b}' w_t + v_t$, where $\bar{a} = (1-\alpha_1)^{-1}\alpha_0$, $\overline{b} = (I - \alpha_1 F')^{-1}\beta$.

Agents are assumed to use the PLM $p_t = a + b' w_t + \eta_t$, where η_t is a disturbance term. The PLM has the same functional form as the REE but possibly different coefficients since agents do not know the REE. To estimate the PLM, agents use data $\{p_i, w_i\}_{i=0}^{t-1}$ and forecast using the estimated model $E^*_t p_{t+1} = a_{t-1} + b'_{t-1} F w_t$.

These forecasts lead to a temporary equilibrium or *actual law of motion* (ALM) $p_t = T(\varphi_{t-1})' z_t + v_t$, where $T(\varphi)' = (\alpha_0 + \alpha_1 a, \alpha_1 b' F + \beta')$. The REE $\left(\bar{a} + \bar{b}'\right)$ is a fixed point of the mapping $T(\varphi)$ from the PLM to the ALM. If we let $\varphi_t' = \left(a_t, b_t'\right)$ and $z_t' = \left(1, w_t'\right)$, RLS estimation is given by

$$\begin{aligned}\varphi_t &= \varphi_{t-1} + t^{-1} R_t^{-1} z_t \left(p_t - \varphi_{t-1}' z_t\right) R_t \\ &= R_{t-1} + t^{-1}\left(z_t z_t' - R_{t-1}\right). \end{aligned} \tag{2}$$

where p_t is given by the ALM. We say that the REE is *stable under RLS learning* if $\left(a_{t-1}, b_{t-1}'\right) \to \left(\bar{a}, \bar{b}'\right)$ over time.

This model of learning involves bounded rationality. Each period agents maximize their objective, given their forecasts. However, agents treat the economy as having constant parameters, which is true only in the REE. Outside the REE the PLMs are misspecified, but misspecification vanishes as learning converges to the REE.

A key result, which holds in numerous models, is that RLS learning converges to RE under certain conditions on model parameters. Thus, the REE can be learned even though economic agents initially have limited knowledge and are boundedly rational.

Expectational stability (*E-stability*) is a convenient way for establishing the convergence conditions for RLS learning. Define the differential equation $d\varphi/d\tau = T(\varphi) - \varphi$, which describes partial adjustment in virtual time τ. The REE is *E-stable* if it is locally stable under the differential equation. For models of the form (1), convergence is guaranteed if $0 < \alpha_1 < 1$, which is satisfied in the Cagan model since $\alpha_1 = \psi(1 + \psi)^{-1}$. Evans and Honkapohja (2001) contains a detailed discussion of convergence of RLS learning.

The Roles of Learning

Adaptive learning has several other important roles besides being a stability theory for REE. RE models can have multiple stationary equilibria, that is, *indeterminacy of equilibrium*. In such situations learning stability acts as a *selection criterion* to determine the plausibility of a particular REE.

As an example consider the non-stochastic Cagan model with government spending financed by seigniorage, with nonlinear reduced form $x_t = G\left(x_{t+1}^e\right)$, where x_t denotes inflation (see Evans and Honkapohja 2001, chs. 11 and 12, for details). This model has two (interior) steady state solutions $\hat{x} = G(\hat{x})$. The low-inflation steady state x_L is stable under learning and the high-inflation steady state x_H is not.

Learning selects a unique REE x_L in this model. In more general models, learning stability does not necessarily select a unique REE, but the set of 'plausible' REE is usually significantly smaller than the set of all REE.

The roles of RLS learning are not restricted to stability of REE and equilibrium selection. Learning can also provide new forms of dynamics as discussed below.

Monetary Policy Design

Indeterminacy of equilibria and instability of REE under RLS learning mean that the economy can be subject to persistent fluctuations. These instabilities can arise in the New Keynesian (NK) model (Woodford 2003), which is widely used for studying monetary policy. Policy design has an important role in eliminating these instabilities and facilitating convergence to 'desirable' equilibria.

Consider the linearized NK model. The IS and PC curves $x_t = -\phi\left(i_t - E_t^* \pi_{t+1}\right) + E_t^* x_{t+1} + g_t$ and $\pi_t = \lambda x_t + \beta E_t^* \pi_{t+1} + u_t$ summarize private sector behaviour. Here x_t, π_t and i_t denote the output gap, inflation and the nominal interest rate. ϕ and λ are positive parameters while $0 < \beta < 1$ is the discount factor. The shocks g_t and u_t are assumed to be observable and follow a known *VAR*(1) process.

Central bank (CB) behaviour is described by an interest-rate rule. CB may use an instrument rule that is not based on explicit optimization. Examples are Taylor rules that depend on current data or forecasts, $i_t = \chi_\pi \pi_t + \chi_x x_t$ or $i_t = \chi_\pi E_t^* \pi_{t+1} + \chi_x E_t^* x_{t+1}$, where $\chi_\pi, \chi_x > 0$.

The IS and PC equations, together with either Taylor rule, lead to a bivariate reduced form in

(x_t,π_t), which can be examined for determinacy (uniqueness of equilibrium) and E-stability. Bullard and Mitra (2002) show that current-data Taylor rules yield both E-stability and determinacy iff $\lambda(\chi_\pi - 1) + (1 - \beta)\chi_x > 0$. Under forward-looking rules $\chi_\pi > 1$ and small χ_x yield E-stability and determinacy.

Optimal monetary policy under discretion and commitment has been examined by Evans and Honkapohja (2003a, b, 2006). Various ways to implement optimal policy have been suggested. Some commonly suggested interest-rate rules, based on fundamental shocks and variables, can lead to E-instability and/or indeterminacy. Evans and Honkapohja advocate appropriate expectations-based rules that deliver both E-stability and determinacy.

Other aspects of learning are also important for monetary policy. One practical concern is the observability of private forecasts needed for forecast-based rules. Results by Honkapohja and Mitra (2005) show that using internal CB forecasts in place of private sector expectations normally delivers E-stability.

Another difficulty for optimal monetary policy is that it requires knowledge of structural parameters, which are in practice unknown. CB can learn the values of φ and λ by estimating IS and PC equations. Expectations-based optimal rules continue to deliver stability under simultaneous learning by private agents and the CB (see Evans and Honkapohja 2003a, 2003b).

Fluctuations

A major issue in macroeconomics is economic fluctuations, for example, business cycles and asset price movements. Can learning help to explain these phenomena?

Stable Sunspot Fluctuations

One theory of macroeconomic fluctuations interprets them as rational 'sunspot' equilibria. Although many macroeconomic models – for example, the real business cycle (RBC) model or Taylor's overlapping contracts model – have a unique stationary solution under RE, other models can have indeterminacy. Examples include the overlapping generations (OLG) model and RBC models with increasing returns and monopolistic competition or tax distortions.

When multiple equilibria are present, some solutions may depend on variables, 'sunspots', that are completely extraneous to the economy. Such stationary sunspot equilibria (SSEs) exhibit self-fulfilling prophecies with the sunspot acting as a coordinating device: if expectations depend on a sunspot variable, then the actual economy, since it depends on expectations, can also depend rationally on the sunspot.

As already noted, learning stability is a selection device. Suppose agents' forecasts are a linear function of both the macroeconomic state and a sunspot variable. If the forecast functions have coefficients close to but not equal to SSE values, and if agents update the estimated coefficients using RLS, can the coefficients converge to SSE values? If not, this casts doubt on the plausibility of SSEs.

SSEs appear not to be stable under learning in indeterminate RBC models but are learnable in some other models. We first describe results for the NK model and then discuss the possibility of stable SSE in other models.

SSEs in the NK Model

Consider again the linearized NK model augmented by either the current-data or forward-looking Taylor rule. As noted above, indeterminacy is likely when the 'Taylor principle' $\chi_\pi > 1$ is violated.

In practice CBs are said to use forward-looking rules, and Clarida et al. (2000) argue that empirical estimates of χ_π are less than 1 in the period before 1984, while they are greater than 1 for the subsequent period. Could SSEs explain the higher economic volatility in the earlier period?

Honkapohja and Mitra (2004) and Evans and McGough (2005) approach this question by asking when SSEs are stable under learning in the NK model. Surprisingly, SSEs appear never to be stable under learning for current-data Taylor rules. When the forward-looking Taylor rule is

employed, stable SSEs occur not when $\chi_\pi < 1$, but rather when $\chi_\pi > 1$ and χ_π and χ_x are sufficiently large, that is, *overly* aggressive rules lead to learnable SSEs. However, this does not rule out the Clarida, Gali, Gertler explanation for pre-1984 instability because, if $\chi_\pi < 1$ leads to indeterminacy, *no* REE is stable under learning and aggregate instability would presumably result.

Stable SSEs in Other Models

Stability under learning is a demanding test for SSEs that is met in only some cases in the NK model. There are, however, other examples of stable SSEs, such as the basic OLG model.

Some nonlinear models can have multiple steady states that are locally stable under RLS learning. In this case there can also be SSEs that take the form of occasional random shifts between neighbourhoods of the distinct stable steady states. Examples of this are the 'animal spirits' model of Howitt and McAfee (1992), based on a positive search externality, and the 'growth cycles' model of Evans et al. (1998) based on monopolistic competition and complementarities between capital goods.

Two stable steady states also play a role in some important policy models. This can arise in a monetary inflation model with a fiscal constraint, developed by Evans et al. (2001), and in the liquidity trap model of Evans and Honkapohja (2005). In these set-ups policy has an important role in eliminating undesirable steady states.

Dynamics with Constant Gain Learning

An alternative route to explaining economic fluctuations is to modify RLS learning so that more recent observations are given a higher weight. A natural way to motivate this is to assume that agents are concerned about the possibility of structural change. In the RLS formula (2) this can be formally accomplished by replacing t^{-1} with a small 'constant gain' $0 < \gamma < 1$, yielding weights that geometrically decline with the age of observations.

This apparently small change leads to 'boundedly rational' fluctuations, with sometimes dramatic effects. Three main phenomena have emerged. First, as shown by Sargent (1999) and Cho et al. (2002), even when there is a unique equilibrium, occasional 'escape paths' can arise with learning dynamics temporarily driving the economy far from the equilibrium. Sargent shows how the reduction of inflation in the 1982–99 period might be due to such an escape path in which policymakers are led to stop attempting to exploit a perceived (but misspecified) Phillips curve trade-off.

Second, in models with multiple steady states, learning dynamics can take the form of periodic shifts between regimes as a result of intrinsic random shocks interacting with learning dynamics. This is seen in the 'increasing social returns' example of Evans and Honkapohja (2001), the hyperinflation model of Marcet and Nicolini (2003), the exchange rate model of Kasa (2004) and the liquidity trap model of Evans and Honkapohja (2005).

Third, even when large escapes do not arise, there can be policy implications, because constant gain learning differs in small but persistent ways from full rationality. Orphanides and Williams (2005) show that policymakers attempting to implement optimal policy should be more hawkish against inflation than under RE.

Other Developments

There continue to be many new applications of learning dynamics in macroeconomics, with closely related work in asset pricing and game theory.

One recent topic concerns the possibility that agents use a misspecified model. Under RLS learning agents may still converge, but to a restricted perceptions equilibrium, rather than to an REE (see Evans and Honkapohja 2001). Another recent development is to allow agents to select from alternative predictors. In the Brock and Hommes (1997) model agents choose, based on recent past performance, between a costly sophisticated and a cheap naive predictor. This can lead to complex nonlinear dynamics. Branch and Evans (2006) combine dynamic predictor selection with RLS learning and show the

existence of 'misspecification equilibria' when all forecasting models are underparameterized.

Other topics and applications include empirical work on expectation formation, calibration and estimation of learning models to data, interaction of policymaker and private-sector learning, learning and robust policy, experimental studies of expectation formation, the role of calculation costs, expectations over long horizons, alternative learning algorithms, expectational and structural heterogeneity, transitional learning dynamics, consistent expectations and near-rationality.

Current interest in learning dynamics is evidenced by five recent Special Issues devoted to learning and bounded rationality, in *Macroeconomic Dynamics* (2003), *Journal of Economic Dynamics and Control* (two in 2005), *Review of Economic Dynamics* (2005), and *Journal of Economic Theory* (2005).

See Also

- Animal Spirits
- Determinacy and Indeterminacy of Equilibria
- Expectations
- Learning and Evolution in Games: An Overview
- Multiple Equilibria in Macroeconomics
- New Keynesian Macroeconomics
- Rational Expectations
- Rationality, Bounded
- Sunspot Equilibrium
- Two-Stage Least Squares and the *k*-Class Estimator

Bibliography

Beck, G., and V. Wieland. 2002. Learning and control in a changing environment. *Journal of Economic Dynamics and Control* 26: 1359–1377.

Branch, W., and G. Evans. 2006. Intrinsic heterogeneity in expectation formation. *Journal of Economic Theory* 127: 264–295.

Brock, W., and C. Hommes. 1997. A rational route to randomness. *Econometrica* 65: 1059–1095.

Bullard, J., and K. Mitra. 2002. Learning monetary policy rules. *Journal of Monetary Economics* 49: 1105–1129.

Cho, I.-K., N. Williams, and T. Sargent. 2002. Escaping Nash inflation. *Review of Economic Studies* 69: 1–40.

Clarida, R., J. Gali, and M. Gertler. 2000. Monetary policy rules and macroeconomic stability: Evidence and some theory. *Quarterly Journal of Economics* 115: 147–180.

Evans, G., and S. Honkapohja. 2001. *Learning and expectations in macroeconomics*. Princeton: Princeton University Press.

Evans, G., and S. Honkapohja. 2003a. Expectations and the stability problem for optimal monetary policies. *Review of Economic Studies* 70: 807–824.

Evans, G., and S. Honkapohja. 2003b. Adaptive learning and monetary policy design. *Journal of Money Credit and Banking* 35: 1045–1072.

Evans, G., and S. Honkapohja. 2005. Policy interaction, expectations and the liquidity trap. *Review of Economic Dynamics* 8: 303–323.

Evans, G., and S. Honkapohja. 2006. Monetary policy, expectations and commitment. *Scandinavian Journal of Economics* 108: 15–38.

Evans, G., S. Honkapohja, and R. Marimon. 2001. Convergence in monetary inflation models with heterogeneous learning rules. *Macroeconomic Dynamics* 5: 1–31.

Evans, G., S. Honkapohja, and P. Romer. 1998. Growth cycles. *American Economic Review* 88: 495–515.

Evans, G., and B. McGough. 2005. Monetary policy, indeterminacy and learning. *Journal of Economic Dynamics and Control* 29: 1809–1840.

Guesnerie, R. 2005. *Assessing rational expectations 2: 'Eductive' stability in economics*. Cambridge, MA: MIT Press.

Honkapohja, S., and K. Mitra. 2004. Are non-fundamental equilibria learnable in models of monetary policy? *Journal of Monetary Economics* 51: 1743–1770.

Honkapohja, S., and K. Mitra. 2005. Performance of monetary policy with internal central bank forecasting. *Journal of Economic Dynamics and Control* 29: 627–658.

Howitt, P., and R. McAfee. 1992. Animal spirits. *American Economic Review* 82: 493–507.

Kasa, K. 2004. Learning, large deviations and recurrent currency crises. *International Economic Review* 45: 141–173.

Marcet, A., and J. Nicolini. 2003. Recurrent hyperinflations and learning. *American Economic Review* 93: 1476–1498.

Orphanides, A., and J. Williams. 2005. Imperfect knowledge, inflation expectations and monetary policy. In *The inflation-targeting debate*, ed. B. Bernanke and M. Woodford. Chicago: University of Chicago Press.

Sargent, T. 1993. *Bounded rationality in macroeconomics*. Oxford: Oxford University Press.

Sargent, T. 1999. *The conquest of American inflation*. Princeton: Princeton University Press.

Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Learning-by-Doing

Spyros Vassilakis

Keywords

Capital–labour ratio; Externalities; Firm size, theory of; Learning spillovers; Learning-by-doing; Partial equilibrium

JEL Classifications

J24

Empirical studies of the production process in various industries have demonstrated a positive association between current labour productivity and measures of past activity like past cumulative output or investment (see Wright 1936; Hirsch 1956; Alchian 1963; Hollander 1965; Sheshinski 1967; Boston Consulting Group 1972, 1974, 1978; Lieberman 1984). A hypothesis advanced to explain this is that labour learns through experience and that experience is obtained during the production process. In other words, learning-by-doing is one of the reasons giving rise to dynamic economies of scale, because a firm knows that increasing current production reduces future average costs. If knowledge obtained within one firm cannot be communicated to other firms, we speak of learning without spillovers. There is some empirical evidence, though, that firms cannot totally exclude outsiders from their stock of knowledge, mainly because of labour turnover (see Boston Consulting Group 1978; Lieberman 1984). Learning spillovers are a special case of positive externalities. The study of learning-by-doing, therefore, is a special case in the study of economies characterized by dynamic economies of scale and positive externalities.

Empirical studies of growth have demonstrated that increases in per capita output cannot be attributed solely, or even mainly, to increases in the capital-labour ratio (see Abramovitz 1956; Solow 1957; Kendrick 1976). On the other hand, Verdoorn (1956) observed a positive relationship between past cumulative output and current labour productivity in the aggregate. This seems to suggest that the part of growth unexplained by increases in the capital-labour ratio could be accounted for by learning-by-doing. Once income per head increases for any reason, say because of an increase in the capital-labour ratio, it will keep on increasing for ever because the initial increase will improve labour productivity and income per head in the next period; after that, the chain of output increases resulting in productivity increases and vice versa is repeated for ever and for the right values of the coefficients it will generate unbounded growth even with stationary population.

Formal models of this process were constructed by Arrow (1962), Levhari (1966), Romer (1986), and Stokey (1986). None of these authors model the process of learning explicitly but consider a world of perfect information with features that are supposed to emerge from a process of learning going on behind the scenes. Here we analyse Romer's model, which is the more general and pays particular attention to existence.

Romer considers a continuous time model with infinitely lived agents who produce and consume a single final good out of a fixed, inexhaustible supply of primary factors. At any given moment in time, the final good can be either consumed or added to the indestructible stock of capital. There is an exogenously given number of firms; each firm's output at any moment in time depends on the amount of capital accumulated by the firm up to that moment, on the amount of natural resources it employs and on the total amount of capital accumulated by all firms up to that moment. In other words, knowledge can be communicated across firms and is incorporated in the capital stock. There are diminishing returns in private capital accumulation but increasing returns when the effect of a firm's accumulation on the total capital stock is taken into account. Notice that the assumption of diminishing returns in private capital accumulation implies that the technical process generated by new learning is capital-augmenting, not land-augmenting. Firms maximize profit and consumers maximize utility taking prices as given. Existence of Walrasian

equilibria is demonstrated under the following assumptions: *(a)* firms do not recognize that their accumulation affects the total capital stock; (*b*) the growth rate of each firm's capital stock is uniformly bounded above and is a concave function of each firm's investment; *(c)* the production function is majorized by a constant plus a constant-elasticity function of the capital stock; (*d*) the discount factor is larger than the product of the above elasticity times the upper bound in the growth rate of the capital stock. Under some additional conditions, the equilibrium capital stock and consumption per head grow without bound. Equilibria are Pareto inefficient because firms do not take into account the fact that their private accumulation adds to the aggregate capital stock and therefore reduces everybody's future costs. Romer and Sasaki (1985) have generated unbounded growth with constant population and a fixed supply of exhaustible resources under more restrictive conditions on the coefficients.

Clearly, the main achievement of Romer's competitive model is to generate unbounded growth without assuming exogenous improvements in technology. The applicability and generality of the model, though, are restricted by the price-taking assumption and the related assumptions that all economies of scale are external to the firm and that the number of firms is fixed. Suppose for a moment that we accept the last two assumptions. Then, as Fudenberg and Tirole (1983) showed, if returns with respect to private capital accumulation are constant, no price-taking equilibrium exists; in other words, the assumption that technical progress is not land-augmenting is crucial. Also, there is no reason why firms should fail to recognize the effect of their actions on the capital stock. Spence (1981) and Fudenberg and Tirole (1983) constructed dynamic partial equilibrium models of learning without spillovers in which a fixed number of firms compete in quantities with Cournot expectations. Industry output may decline over time at a subgame perfect equilibrium, depending on how large is the discount factor relative to the number of firms. Stokey (1986) investigated the same model with spillovers and found that industry output increases over time. Given the importance of spillovers in generating growth, therefore, it seems worthwhile to study their determinants.

The next step is to remove the assumption that all dynamic economies of scale are external to the firm and that the number of firms is fixed. Even if one begins with a situation of purely external economies of scale, there are powerful economic incentives to internalize these economies by reductions in the number of firms, either by collusion or by competition that drives some firms out of business. The tendency to collude to internalize externalities is checked by the incentive of each firm to shirk (underinvest in learning) given that others have done their share of investment. The tendency of competition to reduce the number of firms is checked by entry when the number of firms is so small that a new entrant's gains by wiping out excess profits exceed losses due to lost economies of scale. The learning-by-doing model coupled with such a theory of firm size could generate more predictions about growth, concentration and distribution of income over time.

Finally, one has to address the issue of the evolution of the externalities themselves over time. The extent of learning spillovers is limited by concentration and by the creation of markets in order to transform the external effects into ordinary goods; both of these magnitudes are endogenous, and so the extent of learning spillovers should also be an endogenous variable. The current formulation of learning spillovers assumes a stable, exogenous relationship between measures of past activity and future productivity, but in a long-run model one would not expect such a relationship to hold, exactly because the number of firms and completeness of markets are variable in the long run.

Bibliography

Abramovitz, M. 1956. Resource and output trends in the U.S. since 1870. Occasional paper no. 52, NBER.

Alchian, A. 1963. Reliability of progress curves in airframe production. *Econometrica* 31: 679–693.

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29: 155–173.

Boston Consulting Group. 1972. Perspectives in experience. Technical report.
Boston Consulting Group. 1974. The experience curve reviewed: Price stability. Technical report.
Boston Consulting Group. 1978. Cross sectional experience curves. Technical report.
Fudenberg, D., and J. Tirole. 1983. Learning-by-doing and market performance. *Bell Journal of Economics* 14: 522–530.
Hirsch, W.Z. 1956. Firm progress ratios. *Econometrica* 24 (2): 136–144.
Hollander, S. 1965. *The source of increased efficiency: A study of DuPont rayon plants*. Cambridge, MA: MIT Press.
Kendrick, J.W. 1976. *The formation and stocks of total capital*. New York: NBER.
Levhari, D. 1966. Extensions of Arrow's learning by doing. *Review of Economic Studies* 33: 117–132.
Lieberman, D. 1984. The learning curve and pricing in the chemical processing industries. *RAND Journal of Economics* 15: 213–228.
Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037.
Romer, P., and Sasaki, H. 1985. Monotonically decreasing natural resource prices under perfect foresight. Working paper no. 19, Rochester Center for Economic Research, New York.
Sheshinski, E. 1967. Tests of the learning-by-doing hypothesis. *Review of Economics and Statistics* 49: 568–578.
Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39: 312–320.
Spence, A.M. 1981. The learning curve and competition. *Bell Journal of Economics* 12: 49–70.
Stokey, N. 1986. The dynamics of industry-wide learning. In *Equilibrium analysis: Essays in honor of K.J. Arrow*, ed. W.P. Heller, R.M. Starr, and D.A. Starrett, vol. 2. Cambridge: Cambridge University Press.
Verdoorn, P.J. 1956. Complementarity and long-range projections. *Econometrica* 24: 429–450.
Wright, T.P. 1936. Factors affecting the cost of airplanes. *Journal of Aeronautical Sciences* 3 (4): 122–128.

Least Squares

Halbert White

The method of least squares is a statistical technique used to determine the best linear or nonlinear regression line. The method, developed independently by Legendre (1805), Gauss (1806, 1809) and Adrain (1808), has a rich and lengthy history described in an excellent six-part article by Harter (1974–6). Least squares is the technique most widely used for fitting regression lines because of its computational simplicity and because of particular optimality properties described below. Primary among these are the facts that it gives the best linear unbiased estimator (BLUE) in the case of linear regression and that it gives the maximum likelihood estimator (MLE) in the case of regression with Gaussian (normal) errors.

A regression line is a model of the expectation of a random variable, denoted Y, given a specified set of conditioning variables, denoted X. Y is called the 'dependent' or 'explained' variable, and X is called the vector of 'independent' or 'explanatory' variables. We denote the conditional expectation of Y given X as $E(Y|X)$. A model of this conditional expectation is a function, say f, which depends on the explanatory variables X, and on a vector of parameters, say β, chosen in such a way that for some value of the parameters β^*, $f(X, \beta^*)$ provides the best fitting approximation to $E(Y|X)$.

There are numerous ways in which to measure the goodness of fit of any particular approximation. Perhaps the most important and commonly used criterion is that of mean squared error. The mean squared error of a random variable Z as an approximation to (or estimate of) a random variable Y is defined as

$$\text{mes}\,(Z, Y) = E\left[(Y - Z)^2\right].$$

Here, $(Y - Z)^2$ is the squared error of Z as an approximation to (estimate of) Y. The smaller the mean squared error, the better, because the closer Z is to Y on average. This criterion penalizes an overestimate of Y by the same amount that it penalizes an underestimate of Y of equal magnitude. Of all the functions of X which one might use as an approximation to Y, the best approximation in this sense is given by $E(Y \mid X)$. That is,

$$\text{mes}\,[E(Y|X), Y] \leq \text{mse}\;\; [g(X), Y]$$

for all functions g of X.

Typically, the conditional expectation $E(Y|X)$ is unknown, so one may approximate $E(Y|X)$ using the model $f(X, \beta)$, choosing β^* to satisfy

$$\text{mse}\ [f(X,\beta^*), E(Y|X)] \leq \text{mse}\,[f(X,\beta), E(Y|X)]$$

for all allowable values of β. It can be shown that this holds if and only if

$$\text{mse}\,[f(X,\beta^*), Y] \leq \text{mse}\,[f(X,\beta), Y],$$

for all allowable values of β. Thus, $f(X, \beta^*)$ is equivalently a best approximation to $E(Y|X)$ or a best approximation to Y in this sense.

Because the relevant expectations are usually unknown, it is not possible to find β^* directly by solving the problem

$$\min_{\beta}\ \text{mse}\ [f(X,\beta), Y].$$

Fortunately, the needed expectation can be estimated if one has sample information on Y and X. In economics, this may take the form of time-series, cross-section, or panel (time-series crosssection) observations. Given a sample of n observations on Y and X, denoted (Y_t, X_t), $t = 1, \ldots, n$, it follows generally from the law of large numbers that

$$n^{-1}\sum_{t=1}^{n}[f(X_t,\beta) - Y_t]^2 = E\left\{[f(X,\beta) - Y]^2\right\} + o_{as}(1),$$

where $o_{as}(1)$ denotes terms vanishing with probability one (i.e. almost surely) as n tends to infinity. A useful estimator for β^* can therefore be found by solving the problem

$$\min_{\beta} n^{-1}\sum_{t=1}^{n}[f(X_t,\beta) - Y_t]^2,$$

or the equivalent problem

$$\min_{\beta}\sum_{t=1}^{n}[Y_t - f(X_t,\beta)]^2.$$

This method of estimating β^* is the *method of least squares,* and the resulting estimator is the *least squares estimator,* denoted $\widehat{\beta}_{\text{LS}}$. The quantity $r(Y_t, X_t, \beta) = Y_t - f(X_t, \beta)$ is the 'residual'. The summation above is the 'sum of squared residuals', and it is this sum to which the word 'squares' refers in the phrase 'least squares'. Substituting $\widehat{\beta}_{\text{LS}}$ into r, gives the 'estimated residual', $\widehat{r}_t = r\left(Y_t, X_t, \widehat{\beta}_{\text{LS}}\right)$; substituting $\widehat{\beta}_{\text{LS}}$ A_{S} into $f(X_t, \beta)$ yields the 'fitted value' $\widehat{Y}_t = f\left(X_t, \widehat{\beta}_{\text{LS}}\right)$. Thus, $Y_t = \widehat{Y}_t + \widehat{r}_t$. The quantity $\sum_{t=1}^{n} Y_t^2$ is the 'total sum of squares', the quantity $\sum_{t=1}^{n} \widehat{Y}_t^2$ is the 'explained sum of squares', and the quantity $\sum_{t=1}^{n} \widehat{r}_t^2$ is the 'unexplained sum of squares'.

The properties of $\widehat{\beta}_{\text{LS}}$ are of particular importance; these depend crucially on the properties of Y_t, X_t, and the function $f(X, \beta)$. White (1981) studies the properties of $\widehat{\beta}_{\text{LS}}$ under the following assumptions:

(A1) $\{Y_t, X_t\}$ is a sequence of independent identically distributed (i.i.d.) random variables;

(A2) f: $\mathbb{R}^k \times B \to \mathbb{R}$ is a measurable function on $\mathbb{R}^k$, $k \in N = \{1, 2, \ldots\}$, for each β in B, a compact subset of $\mathbb{R}^p$, $p \in$ N, and a continuous function on B for each x in $\mathbb{R}^k$;

(A3) $q(Y_t, X_t, \beta) = [Y_t - f(X_t, \beta)]^2$ is dominated by an integrable function, i.e., there exists d: $\mathbb{R}^{k+1} \to \mathbb{R}$ such that for all (y, x) in $\mathbb{R}^{k+1}$ and β in B, $q(y, x, \beta) \leq d(y,$ x) and $E[d(Y_t, X_t)] < \infty$;

(A4) $E([Y_t - f(X_t, \beta)]^2)$ has a unique minimum at β^* in B. With these conditions, White proves the following result.

Theorem 1 Given A1–A4, $\widehat{\beta}_{\text{LS}} = \beta * + o_{as}(1)$.

Thus, the least squares estimator $\widehat{\beta}_{\text{LS}}$ converges almost surely to β^*, the parameter value such that $f(X_t, \beta^*)$ provides the minimum mse approximation to Y_t and $E(Y_t \mid X_t)$.

An approximate sampling distribution for $\widehat{\beta}_{\text{LS}}$ exists using the following conditions.

(A4′) A4 holds, and β^* is interior to B;

(A5) $f(x,\cdot)$ is continuously differentiable of order two on B for each x in $\mathbb{R}^k$;

L

(A6) The elements of $\nabla q(Y_t, X_t, \beta)'\nabla q(Y_t, X_t, \beta)$ and $\nabla^2 q(Y_t, X_t, \beta)$ are dominated by integrable functions, where ∇ denotes the gradient operator with respect to β;

(A7) The matrices $A^* = \mathrm{E}[\nabla^2 q(Y_t, X_t, \beta^*)]$ and $B^* = E[\nabla q(Y_t, X_t, \beta^*)' \nabla q (Y_t, X_t, \beta^*)]$ are positive definite.

The result is

Theorem 2 Given

$$\mathrm{A1}-\mathrm{A4}', \ \mathrm{A5}-\mathrm{A7}, \quad n^{1/2}\left(\widehat{\beta}_{\mathrm{LS}} - \beta^*\right) \xrightarrow{d} N\left(0, \ A^{*-1}B^*A^{*-1}\right).$$

Further,

$$\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} = A^{*-1}B^*A^{*-1} + o_{as}(1),$$

where

$$\widehat{A} = n^{-1}\sum_{t=1}^{n} \nabla^2 q\left(Y_t, X_t, \ \widehat{\beta}_{\mathrm{LS}}\right),$$

$$\widehat{B} = n^{-1}\sum_{t=1}^{n} \nabla q\left(Y_t, X_t, \ \widehat{\beta}_{\mathrm{LS}}\right)' \nabla q\left(Y_t, X_t, \ \widehat{\beta}_{\mathrm{LS}}\right).$$

This implies that to test H_0: $s(\beta^*) = 0$ vs. H_a: $s(\beta^*) \neq 0$, where s is a $\upsilon \times 1$ vector function, one can form the Wald statistic

$$W = ns\left(\widehat{\beta}_{\mathrm{LS}}\right)'\left[\nabla\left(\widehat{\beta}_{\mathrm{LS}}\right)\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}\nabla_s\left(\widehat{\beta}_{\mathrm{LS}}\right)'\right]^{-1} s\left(\widehat{\beta}_{\mathrm{LS}}\right)$$

which has the χ^2_υ distribution approximately in large samples under H_0.

An important special case arises when

$$f(X_t, \beta) = \beta_1 + X_t\beta_2,$$

where $\beta' = \left(\beta_1, \ \beta_2'\right)$, with β_1 a scalar and β_2 a $k \times 1$ vector. This is the '(standard) linear model'. In this case,

$$\widehat{\beta}_{\mathrm{LS}} = (X' \ X)^{-1}X'Y,$$

where X is the $n \times k + 1$ matrix with rows $(1, X_t)$, and Y is the $n \times 1$ vector with elements Y_t. This form for $\widehat{\beta}_{\mathrm{LS}}$ is called the 'ordinary least squares estimator'.

Results for the linear model similar to Theorems 1 and 2 above follow by retaining (A1) and imposing

(A2′) $f(x, \ \beta) = \beta_1 + x \ \beta_2$, for $x \in \mathbb{R}^k$, $\beta' = \left(\beta_1, \beta_2'\right) \in \mathbb{R}^{k+1}$;

(A3′) $E\left(Y_t^2\right) < \infty, \ E\left(X_t X_t'\right) < \infty$;

(A4″) $\det E\, E\left(X_t' \ X_t\right) > 0$.

White (1980a) proves the following result.

Theorem 3 Given A1, A2′, A3′, and A4″,

$$\widehat{\beta}_{\mathrm{LS}} = \beta * + o_{as}(1), \text{where } \beta* = \left[E\left(X_t'X_t\right)\right]^{-1}E\left(X_t' \ Y_t\right) < \infty.$$

An asymptotic normality result holds, using the conditions

(A3″) $E\left(Y_t^4\right) < \infty$, and $E\left[\left(X_tX_t'\right)^2\right] < \infty$;

(A5′) $\det E\left(X_t' \ r_t^* r_t^{*'} X_t\right) > 0$, where $r_t^* = r(Y_t, \ X_t, \ \beta^*)$.

The result is

Theorem 4 Given A1, A2′, A3″, A4″, and A5′,

$$n^{1/2}\left(\widehat{\beta}_{\mathrm{LS}} - \beta^*\right) \xrightarrow{d} N\left(0, \ A^{*-1}B^*A^{*-1}\right),$$

where

$$A^* = E\left(X_t' \ X_t\right)$$

and

$$B^* = E\left(X_t' \ r_t^* r_t^{*'} X_t\right).$$

Further,

$$\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} = A^{*-1}B^*A^{*-1} + o_{as}(1),$$

with

$$\widehat{A} = X'X/n, \text{and } \widehat{B} = X'\widehat{\Omega}X/n,$$

where $\widehat{\Omega}$ is the $n \times n$ diagonal matrix with diagonal elements $\widehat{r}_t^2$.

To test the linear hypothesis H_0: $R\beta^* = r$ vs H_a: $R\beta^* \neq r$, where R is a given $\upsilon \times k+1$ matrix and r is a given $\upsilon \times 1$ vector, one can compute

$$W = n\left(R\widehat{\beta}_{\mathrm{LS}} - r\right)'\left[R(X'X|n)^{-1}X'\widehat{\Omega}X/n \times (X'X/n)^{-1}R'\right]^{-1}\left(R\widehat{\beta}_{\mathrm{LS}} - r\right).$$

Under H_0, this has the χ^2_υ distribution approximately in large samples.

Similar results hold in situations more general than the case of i.i.d. observations. See e.g. Domowitz and White (1982) and Gallant and White (1987).

In applications, it is often assumed (with or without justification) that the model $f(X, \beta)$ is correctly specified; that is, there exists β^0 in B such that

$$E(Y|X) = f\left(X, \beta^0\right) \text{ a.s.}$$

For discussion of nonlinear least squares estimation in this context, the reader is referred to Jennrich (1969), Hannan (1971), Klimko and Nelson (1978), White (1980b) and White and Domowitz (1984).

Because of the popularity of the linear model, we discuss the properties of $\widehat{\beta}_{\mathrm{LS}}$ for the correctly specified case in more detail. We adopt the following assumptions.

(B1) $\{Y_t, X_t\}$ is a sequence of independently distributed random variables such that

(a) $E(Y_t) < \infty$, and there exists $\beta^0 \in \mathbb{R}^{k+1}$, $\beta^{0'} = \left(\beta_1^0, \beta_2^{0'}\right)$ such that

$$E(Y_t|X_t) = \beta_1^0 + X_t\beta_2^0, \quad t = 1, 2, \ldots;$$

(b) For all $t = 1, 2, \ldots, E\left(Y_t^4\right) < \infty, E\left[\left(X_tX_t'\right)^2\right] < \infty$, and

$$\mathrm{var}(Y_t|X_t) = \sigma_0^2 \neq 0.$$

(B2) $f(x, \beta) = \beta_1 + x\beta_2$, for $x \in \mathbb{R}^k$, $\beta' = (\beta_1, \beta_2') \in \mathbb{R}^{k+1}$;

(B3) There exist

$$\delta > 0 \text{ and } \Delta < \infty$$

such that

$$E\left|Y_t^2\right|^{1+\delta} < \Delta \text{ and } E\left|X_tX_t'\right|^{1+\delta} < \Delta, t = 1, 2, \ldots;$$

(B4) There exists $\delta > 0$ such that for all n sufficiently large $\det E(X'X/) > \delta$

With these conditions we have

Theorem 5 Given B1a, and B2–B4,

$$\widehat{\beta}_{\mathrm{LS}} = \beta^0 + o_{as}(1)$$

If (B1(b)) also holds, then

$$\boldsymbol{n}^{1/2}\left(\widehat{\beta}_{\mathrm{LS}} - \beta^0\right) \xrightarrow{d} \boldsymbol{N}?(0, \sigma_0^2[\boldsymbol{E}(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n})]^{-1}).$$

Further, $\widehat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n})^{-1} = \sigma_0^2[\boldsymbol{E}(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n})]^{-1} + \boldsymbol{o_{as}}(1)$

To test H_0:$R\beta^0 = r$ versus H_0:$R\beta^0 \neq r$ compute

$$\boldsymbol{W} = \boldsymbol{n}\left(\boldsymbol{R}\widehat{\beta}_{\mathrm{LS}} - \boldsymbol{r}\right)'\left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n})^{-1}\boldsymbol{R}^{-1}\right] \times \left(\boldsymbol{R}\widehat{\beta}_{\mathrm{LS}} - \boldsymbol{r}\right)\widehat{\sigma}^2.$$

Under H_0, this has the χ^2_υ distribution approximately in large samples.

When (B1(b)) is not available, asymptotic normality results for $\widehat{\beta}_{\mathrm{LS}}$ may still hold. We impose

(B3′) There exist $\delta > 0$ and $\Delta < \infty$ such that $E\left|Y_t^4\right|^{1+\delta} < \Delta$ and $E\left|\left(X_tX_t'\right)^2\right|^{1+\delta} < \Delta$, for all $t = 1,2,\ldots$;

(B5) There exists $\delta > 0$ such that for all n sufficiently large,

$$\det n^{-1}\sum_{t-1}^{n} E\left(X_t' \in_t \in_t' X_t\right) > \delta,$$

L

where

$$\in_t = Y_t - E(Y_t|X_t).$$

White (1980c) proves the following result.

Theorem 6 Given B1a, B2, B3′, B4, and B5,

$$\boldsymbol{n}^{1\backslash 2}\left(\widehat{\beta}_{\mathrm{LS}} - \beta^0\right) \xrightarrow{d} \boldsymbol{N}\left(0, \boldsymbol{A}^{0-1}\boldsymbol{B}^0\boldsymbol{A}^{0-1}\right),$$

with

$$\boldsymbol{A}^0 = \boldsymbol{E}(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n}) \text{ and } \boldsymbol{B}^0 = \boldsymbol{E}(\boldsymbol{X}'\Omega\boldsymbol{X}/\boldsymbol{n})$$

where Ω is the $n \times n$ diagonal matrix with diagonal elements ε_t^2.

Further,

$$\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^{-1} = \boldsymbol{A}^{0-1}\boldsymbol{B}^0\boldsymbol{A}^{0-1} + \boldsymbol{o}_{\boldsymbol{as}}(1)$$

with

$$\widehat{\boldsymbol{A}} = \boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n} \text{ and } \widehat{\boldsymbol{B}} = \boldsymbol{X}'\widehat{\Omega}\boldsymbol{X}/\boldsymbol{n},$$

where $\widehat{\Omega}$ is as previously defined.

This result allows hypothesis testing even when the 'errors' ε_t exhibit heteroskedasticity of unknown form. To test H_0: $R\beta^0 = \mathrm{r}$ versus H_a: $R\beta^0 \neq \mathrm{r}$, compute

$$\boldsymbol{W} = \boldsymbol{n}\left(\boldsymbol{R}\widehat{\beta}_{\mathrm{LS}} - \boldsymbol{r}\right)' \times \left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n})^{-1}\boldsymbol{X}'\widehat{\Omega}\boldsymbol{X}/\boldsymbol{n}(\boldsymbol{X}'\boldsymbol{X}/\boldsymbol{n})^{-1}\boldsymbol{R}'\right]^{-1}\left(\boldsymbol{R}\widehat{\beta}_{\mathrm{LS}} - \boldsymbol{r}\right).$$

Under H_0, this has the χ_v^2 distribution approximately in large samples.

Under slightly different assumptions, the properties of $\widehat{\beta}_{\mathrm{LS}}$ can be specified for samples of any size. Retaining (B1) and (B2), we replace (B3) and (B4).

(B3″) For all $t = 1, 2, \ldots, E\left(Y_2^t\right) < \infty$ and $E\left(X_t X_t'\right) < \infty$;

(B4′) For any $n \geq k + 1$, $P[\det X'X > 0] = 1$.

Now we have

Theorem 7 Given B1(a), B2, B3″, and B4′, for any $n \geq k + 1$,

$$\boldsymbol{E}\left(\widehat{\beta}_{\mathrm{LS}}|\boldsymbol{X}\right) = \beta^0 \text{a.s.}$$

so that $\boldsymbol{E}\left(\widehat{\beta}_{\mathrm{LS}}\right) = \beta^0$.

That is, $\widehat{\beta}_{\mathrm{LS}}$ is unbiased, conditionally and unconditionally.

Modifying B1, we obtain the sampling distribution for $\widehat{\beta}_{\mathrm{LS}}$ in samples of any size.

(B1′) $\{Y_t, X_t\}$ is a sequence of independent random variables such that $E\left(Y_t^2\right) < \infty$, and for some $\beta^{0'} = \left(\beta_2^0, \beta_2^{0'}\right)$ in $\mathbb{R}^{k+1}$,

$$Y_t|X_t \sim N\left(\beta_1^0 + X_t\beta_2^0, \sigma_0^2\right), 0 < \sigma_0^2 < \infty, \quad t = 1, 2, \ldots$$

We have the following version of the 'classical' sampling distribution theorem for the least squares estimator.

Theorem 8 Given B1′, B2, B3″, and B4′, for any $n \geq k + 1$,

$$\widehat{\beta}_{\mathrm{LS}}|\boldsymbol{X} \sim \boldsymbol{N}\left(\beta^0, \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

and

$$(\boldsymbol{n} - \boldsymbol{k} - 1) - \widehat{\sigma}_{\mathrm{LS}}^2/\sigma_0^2|\boldsymbol{X} \sim \chi_{\boldsymbol{n}-\boldsymbol{k}-1}^2,$$

where $\widehat{\sigma}_{\mathrm{LS}}^2 = \left(\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right)/(\boldsymbol{n} - \boldsymbol{k} - 1)$ is independent of $\widehat{\beta}_{\mathrm{LS}}$ conditional on X.

To test the linear hypothesis H_0: $R\beta^0 = \mathrm{r}$ vs H_a: $R\beta^0 \neq \mathrm{r}$ one can use Fisher's F-statistic

$$\frac{F\left(\boldsymbol{R}\widehat{\beta}_{\mathrm{LS}} - \boldsymbol{r}\right)'\left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1}\left(\boldsymbol{R}\widehat{\beta}_{\mathrm{LS}} - \boldsymbol{r}\right)/v}{\widehat{\sigma}_{\mathrm{LS}}^2/(\boldsymbol{n} - \boldsymbol{k} - 1)}$$

Under H_0, this has Fisher's F-distribution with v, $n - k - 1$ degrees of freedom. When $v = 1$, another statistic is available. Given a $1 \times k + 1$ weighting vector w, and a scalar w_0, one can test H_0: $w\beta^0 = w_0$ vs, H_a: $w\beta^0 \neq w_0$, using Student's t-statistic

$$t = \left(w\widehat{\beta}_{\mathrm{LS}} - w_0 \right) / \left[\widehat{\sigma}^2_{\mathrm{LS}} w' (X'X)^{-1} w \right]^{1/2}.$$

Under H_0, this has Student's t-distribution with $n - k - 1$ degrees of freedom.

Finally, the least squares estimator for the correctly specified linear model has desirable efficiency properties. The Gauss–Markov Theorem states that the ordinary least squares estimator is the best linear unbiased estimator (BLUE) in the sense that any other estimator constructed as a linear combination of Y – say $W'Y$, where W is a $k + 1 \times n$ matrix depending only on X – which is unbiased (i.e. $E(WY) = \beta^0$), has a variance–covariance matrix which differs from that of $\widehat{\beta}_{\mathrm{LS}}$ by a positive semi-definite matrix. This holds whether or not Y_t has a normal distribution conditional on X_t. When Y_t does have the normal distribution conditional on X_t, $\widehat{\beta}_{\mathrm{LS}}$ is the maximum likelihood estimator (MLE) and is therefore the best unbiased estimator, as the MLE attains the Cramer–Rao bound, conditional on X. For a more detailed discussion of the least squares estimator in this context, see Theil (1971), Johnston (1984), and White (1984).

See Also

- ▶ Econometrics
- ▶ Estimation
- ▶ Maximum Likelihood

Bibliography

Adrain, R. 1808. Research concerning the probabilities of errors which happen in making observations. *Analyst* 1: 93–109.

Domowitz, I., and H. White. 1982. Misspecified models with dependent observations. *Journal of Econometrics* 20: 35–58.

Gallant, A.R., and H. White. 1987. *A unified theory of estimation and inference for nonlinear dynamic models*. Oxford: Basil Blackwell.

Gauss, C.F. 1806. II Comet vom Jahr 1805. *Monatliche Correspondenz zur Beförderung der Erd- und Himmelskunde* 14: 181–186.

Gauss, C.F. 1809. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Hamburg: F. Perthes and I.H. Besser.

Hannan, E.J. 1971. Nonlinear time-series regression. *Journal of Applied Probability* 8: 767–780.

Harter, H.L. 1974–6. The method of least squares and some alternatives, Parts I–VI. *International Statistical Review* 42: 147–174, 235–264, 282; 43, 1–44, 125–190, 269–278; 44, 113–159.

Jennrich, R. 1969. Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* 40: 633–643.

Johnston, J. 1984. *Econometric methods*. New York: McGraw-Hill.

Klimko, L., and P. Nelson. 1978. On conditional least squares estimation for stochastic processes. *Annals of Statistics* 6: 629–642.

Legendre, A.M. 1805. *Nouvelles méthodes pour la determination des orbites des comètes*. Paris: Courcier.

Theil, H. 1971. *Principles of econometrics*. New York: Wiley & Sons.

White, H. 1980a. Using least squares to approximate unknown regression functions. *International Economic Review* 21: 149–170.

White, H. 1980b. Nonlinear regression on cross-section data. *Econometrica* 48: 721–746.

White, H. 1980c. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838.

White, H. 1981. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76: 419–433.

White, H. 1984. *Asymptotic theory for econometricians*. Orlando: Academic Press.

White, H., and I. Domowitz. 1984. Nonlinear regression with dependent observations. *Econometrica* 52: 143–162.

Lederer, Emil (1882–1939)

Robert A. Dickler

Keywords

Austrian economics; Bureaucracy; Business cycles; Cartels; Class; Creative destruction; Division of labour; Entrepreneurship; Excess capacity; Great Depression; Increasing returns; Lederer, E.; Mixed economy; Planning; Structural unemployment; Technical change; Uncertainty; Vertical integration

JEL Classifications

B31

Emil Lederer was a prominent economist and sociologist in the German Weimar Republic. He was born in the Bohemian town of Pilsen in 1882 and died in political exile in New York City in 1939. In Vienna and Berlin, where he studied both law and economics, Lederer participated in advanced seminars conducted by Menger, Böhm-Bawerk and Schmoller. From 1918 to 1931 he served as professor in Heidelberg and then succeeded Sombart in Berlin from 1931 to 1933. In collaboration with E. Jaffé, Schumpeter and Sombart as well as Max and Alfred Weber, he edited the *Archiv für Sozialwissenschaft und Sozialpolitik,* the renowned social science journal which ceased publication under the Nazi regime. After emigrating to New York in 1933 be became the first Dean of the New School for Social Research's Faculty of Political and Social Sciences, which was comprised of outstanding Continental scholars who had also sought asylum in the United States.

Lederer made pioneering contributions towards understanding the social, political and economic significance of large-scale, bureaucratic private enterprise. In a major theoretical and empirical study based on his *Habilitation*, Lederer undertook the first comprehensive analysis of the working conditions and political attitudes of salaried employees (Lederer 1912). Subsequent work together with Jacob Marschak showed how rationalization of production along with bureaucratic division of labour in administration formed the basis for the rise of the new middle class (Lederer and Marschak 1926). They concluded that the evolution of class structure in advanced capitalist societies undermines political stability and raises the spectre of fascism. Anxiety stemming from economic insecurity and abhorrence of collective action with organized labour weakens the growing middle class's support for democratic forms of government and strengthens its tolerance of authoritarian institutions to suppress the demands of the proletariat.

Lederer's advanced economics textbook contains an authoritative exposition and critique of objective and subjective value theories (Lederer 1931a). The laws of the market economy, as depicted in the marginalist doctrine, are no longer in effect, since economies of large-scale production prevail. Adoption of modern technologies requires vertical integration, high proportion of fixed capital and substantial fixed costs for sales and general administrative overhead. Complementarities and decreasing marginal costs are the rule in basic industries (coal, steel, chemicals and utilities).

Increasing returns to scale is not only an anomaly which cannot be subsumed under the marginalist paradigm; it forms the starting point for business cycle theory (Lederer 1924, 1927). Disproportionalities in growth of demand for investment and consumer goods are due to unavoidable price inflexibilities and absence of strong equilibrating tendencies. Cartels which administer prices and set production quotas are the natural outcome of the technically determined drive to realize economies of scale. The self-contained planning of separate industrial bureaucracies lacks interindustry coordination and thus cannot prevent misallocation, underutilization and periodic decumulation of capital.

Rapid labour-saving technical change is regarded by Lederer as a key factor in explaining the severity of unemployment during the Great Depression (Lederer 1931b, 1936a). In an upswing, dynamic enterprises exploit opportunities to realize above-normal returns on investment offered by introduction of highly mechanized techniques. Labour is displaced not only by rationalization of operations but also by diversion of capital from static enterprises which do not employ the new techniques. As productivity and productive capacity in dynamic enterprises increases, monopolistic market structures prevent prices from falling faster than wages. Redistribution of income from labour to capital decreases consumer goods demand, which in turn reduces the derived demand for capital goods and brings about excess capacity in capital goods production. Without incentives for accelerating the form of technical progress which creates new products, opens up new markets and stimulates labour-absorbing investment, technological unemployment persists.

Stressing the distinction between labour-saving and labour-absorbing forms of technical

progress, Lederer criticized Keynes for his failure to analyse long-run dynamics (Lederer 1936b). Investment in plant and equipment embodies new techniques. Not the lack of profitable investments, but rather an abundance of abnormally profitable rationalization investments creates structural unemployment in addition to the cyclical unemployment treated by Keynes. Government spending is necessary to stimulate the economy, but it is not sufficient to overcome mass unemployment. Democratic national planning is also necessary to attract capital to new industries offering additional employment opportunities.

Lederer's conviction that a mix of market and planned economies based on political consensus is practicable may be traced to his close association with the industrialist and statesman, Walther Rathenau, who was the architect of German economic mobilization in the First World War (Lederer 1933, 1934).

Along with Schumpeter, Lederer cultivated an undogmatic Austrian style of theorizing. Both emphasized the significance of uncertainty, entrepreneurship (or its absence as a consequence of bureaucratization), disequilibrating forces, such as technical change, and underlying instability of capitalism. Schumpeter (1939, 1942) defended neoclassical equilibrium theory by asserting that the price system it represents moves automatically, but not without friction, towards a new equilibrium following the 'creative destruction' of an old equilibrium. Similarly, Lederer (1931b) wrote: 'The capitalist dynamic is not only "development" but also "destruction" '. However, Lederer combined neo-Ricardian (von Bortkiewicz 1907) and Austro- Marxian (Hilferding 1910) approaches to focus on the production system; accordingly, in his view there was no automatic mechanism to assure that investment brings about a rate and direction of technical change consistent with full employment equilibrium.

Selected Works

1912. *The problem of the modern salaried employee: Its theoretical and statistical basis.* Trans. E.E. Warburg. New York: State Department of Social Welfare and the Department of Social Science, Columbia University, 1937.

1924. Konjunktur und Krise. *Grundriss der Sozialökonomik* 4(1). Tübingen: J.C.B. Möhr.

1926. (With J. Marschak.) *The new middle class.* Trans. S. Ellison. New York: State Department of Social Welfare and the Department of Social Science, Columbia University, 1937.

1927. Monopol und Konjunktur. *Vierteljahrshefte zur Konjunkturforschung.* Vol. 2, Supplement 2. Berlin: Duncker & Humblot.

1931a. *Aufriss der ökonomischen Theorie.* 3rd edn. Tübingen: J.C.B. Möhr.

1931b. *Technischer Fortschritt und Arbeitslosigkeit.* Tübingen: J.C.B. Möhr.

1933. National economic planning. In *Encyclopaedia of the social sciences.* Vol. 11. New York: Macmillan.

1934. Rathenau, Walther. In *Encyclopaedia of the social sciences.* Vol. 13. New York: Macmillan.

1936a. *Technical progress and unemployment. An enquiry into the obstacles to economic expansion*. Studies and reports, Series C, No. 22. Geneva: International Labor Office. Trans. of revised edition of (1931a).

1936b. Commentary on Keynes. *Social Research* 3, 478–87.

1979. *Kapitalismus, Klassenstruktur und Probleme der Demokratie in Deutschland 1910–1940.* Göttingen: Vandenhoeck & Ruprecht. (Collection of essays, edited by J. Kocka with biographical essay by Hans Speier.)

Bibliography

Bortkiewicz, L. von. 1907. On the correction of Marx's fundamental theoretical construction in the third volume of *Capital*. In *Karl Marx and the close of his system by E. von Böhm-Bawerk and Böhm-Bawerk's Criticism of Marx by R. Hilferding*, ed. P.M. Sweezy. New York: Kelley, 1949.

Hilferding, R. 1910. *Das Finanzkapital*. Vienna: I. Brand.

Schumpeter, J.A. 1939. *Business cycles*. New York: McGraw-Hill.

Schumpeter, J.A. 1942. *Capitalism, socialism and democracy.* 2nd edn,. New York: Harper, 1947.

Lefebvre, Georges (1874–1959)

Robert Forster

Lefebvre was one of the most prolific and influential French historians of the first half of the 20th century. This is especially striking since he published his first book at the age of fifty. His name has become almost interchangeable with the history of the French Revolution. Lefebvre's work falls into two broad categories: in-depth studies of the French peasantry based on exhaustive archival research, and three massive synthetic works on the French Revolution and Napoleon which have been unsurpassed in their thoroughness and objectivity.

More of a social and political historian than an economic one, Lefebvre defies easy classification, while the solidity of his research has led more than one school of history to claim him as their own. Jacobin–Marxist historians, for example, have placed him in a hallowed historical tradition reaching back to Jules Michelet and Jean Jaurès in the 19th century and made him a vital link in a longer chain of French Socialist and Communist historians who regard the French Revolution as the first step toward a future 'classless society'. Although his name was prominently displayed on the cover of the *Annales Historiques de la Révolution Française* from 1932 to 1959, Lefebvre shared little of the doctrinaire certainty of Albert Mathiez (1874–1932) or Albert Soboul (1914–1982), his colleagues and fellow historians of the French Revolution. A socialist by political persuasion and humanitarian inclination, Lefebvre did not interpret the French Revolution primarily as a clash of material class interests or productive forces. In his famous essay, *Quatre-vingt-neuf* (translated by R.R. Palmer as *The Coming of the French Revolution*, 1947), Lefebvre refused to reduce the bourgeoisie to a paradigm of Balzacian greed and stressed the universal and beneficial applicability of the ideology of the Rights of Man, however incomplete the realization of equality.

Lefebvre never lost sight of individuals, especially ordinary humble individuals, and though he employed the terminology of class, he was always aware of social hybrids, multiple economic roles, and the intractable ambiguities of group behaviour and attitudes. This breadth of vision and respect for nuance was especially apparent in his work on the French peasantry in which a knot of capitalist farmers (*gros fermiers*) were invariably surrounded by a mass of micro-owners who combined the economic functions of day-labourers, sharecroppers, tenant farmers, quit-renters (*censitaires*), and proprietors. Coherent and uniform interests, attitudes, and goals could not easily be deduced from such functional pluralism. Lefebvre knew peasant society in all of its complexity. He was convinced that no simple bi-polar class struggle could adequately explain it.

Lefebvre's work represents a meeting place of two major schools of French historiography: The *longue durée* of rural history (see M. Bloch and F. Braudel and the *rupture* so celebrated by both Marxist and Whig historians of the French Revolution. Lefebvre attempted to resolve this tension by emphasizing the limitations of the Revolution (and by inference of any national revolution with a large urban component) for the peasants. Although benefitting from the abolition of the seigneurial system and of a society of privileged orders, they failed to obtain any significant redistribution of the land or even greater tenant security. Lefebvre attributed this failure not only to the ideology of the bourgeois leadership, but also to the conflicting interests and values of the peasants themselves. However, viewed from a national and even international perspective, the French Revolution effected a major change, not primarily in class alignments, but in the confirmation of new political ideas and institutions. The foundation of a new society had been laid, 'new' in its values, not in its class base. The Declaration of the Rights of Man and the Citizen was its central message, a message

applicable to all human kind and not only to a minority of owners of capital.

If the subject-matter of Lefebvre's work falls into two broad categories – peasant studies and the French Revolution – his approach to history followed at least three paths: social structure, political change, and collective psychology. Lefebvre's study of mass behaviour, especially under crisis conditions, was one of his most original contributions to historical studies. *La Grande Peur de 1789* is a classic of this type. Written in 1932, the book anticipated by almost a half-century a large historical literature on crowd behaviour. Lefebvre described in immediate personal terms fear, panic, and rumour; he charted their relays, circuits, and warning points and demonstrated their consequences for the revolutionary crisis of 1789. The 'aristocratic plot', which Lefebvre identified as largely a figment of peasant imagination, became more important to understanding mass actions than the formal policies of government, the price of bread, or the privileges of the elites. He demonstrated that it is not enough to list 'causal factors'; the historian must chart their translation into behaviour. A large place must be made for the irrational, the unexpected, and the contingent. Human beings cannot be adequately explained or categorized by their occupation, revenue, residence, or *état civil*, and surely not by rational goals postulated by social scientists. This was especially evident in moments of social crisis.

Georges Lefebvre was more than the author of ten major works of history, editor of the *Annales Historiques de la Révolution Française*, professor at the Sorbonne, and holder of the Chair in the French Revolution from 1937 to 1945. He was also a great teacher and a warm and generous person. One day in the autumn of 1953 a young student left Lefebvre's modest house in the outskirts of Paris only to hear a cry over his shoulder. Professor Lefebvre was in the street, clad in his slippers, cupping his hands to his mouth: 'Forster, n'oubliez pas la série E aux archives de Toulouse!' He was 79 years old.

Selected Works

1924. *Les paysans du Nord pendant la révolution française*. Paris/Lille: F. Rieder & Co.

1931. *La révolution française*. Paris: F. Alcan. 2nd ed., Paris: Presses Universitaires, 1951. Trans. by Elizabeth Moss Evanson as *The French revolution*. London: Routledge & Kegan Paul, 1962–4.

1932a. *La grande peur de 1789*. Paris: A. Colin. 2nd ed., Paris: Libraire Armand Colin, 1970. Trans. by Joan White as *The great fear of 1789*. London: New Left Books, 1973.

1932b. *Questions agraires au temps de la Terreur*. Paris: Commission d'Histoire économique de la Révolution. 2nd ed., La Roche-sur-Yon: H. Potier, 1954.

1935. *Napoléon. Paris: Presses Universitaires*. 4th ed., 1952. English trans., London: Routledge & Kegan Paul, 1973.

1937. *Les Thermidoriens*. Paris: A. Colin. Trans. by Robert Baldrick as *The Thermidoreans*. New York: Random House; London: Routledge & Kegan Paul, 1964.

1939. *Quatre-vingt-neuf*. Paris: Maison Livre Francais. Trans. by R.R. Palmer as *The coming of the French revolution*. Princeton: Princeton University Press, 1947. Reprinted 1967.

1946. *Le Directoire*. Paris: A. Colin. Trans. by Robert Baldrick as *The directory*. New York: Random House; London: Routledge & Kegan Paul, 1964.

1962–3. *Etudes Orléanaises,* 2 vols. Paris: Commission d'Histoire économiques et social de la Révolution.

1965. *Cherbourg à la fin de l'ancien régime et au début de la révolution*. Caen. Collections of sources Edited by Lefebvre

1914, 1921. *Documents relatifs à l'histoire des subsistances dans le district de Bergues 1789– an V*, 2 vols. Lille: C. Robbe.

1953. *Recueil de documents relatifs aux séances des Etats-Généraux de 1789*. Paris: Renseignements et vente au Service des publications du Centre national de la recherche scientifique. Translations by Lefebvre

Stubbs, W. 1907, 1923, 1927. *Histoire constitutionelle de l'Angleterre*, 3 vols. Paris: V. Giard & E. Brière.

Legal Institutions and the Ancient Economy

Dennis P. Kehoe

Abstract

This brief survey suggests some of the issues that can be investigated by a careful analysis of the relationship between legal institutions and the economy in the ancient world. By investigating legal institutions, we can better understand the relationships that shaped the economy and the likely implications of these relationships for economic performance. It covers institutions in the Ancient Greek world, in the Ancient Roman world, and more briefly in Ptolemaic Egypt.

Keywords

Agency; Ancient Greece; Ancient Rome; Ancient Egypt; Law and economics; New institutional economics

JEL Classifications

B11

Within the broad constraints imposed by population and technology (Scheidel et al. 2007), law and legal institutions played an important role in ancient economies. The overriding question concerns how formal institutions, including courts and contractual types, and informal ones, such as social conventions or ideology, affected incentives to enter into mutually beneficial contractual arrangements. The alternative is that the laws and legal institutions surrounding an ancient economy served primarily to protect the privileges or interests of certain well-connected groups. Understanding the role of legal institutions in an ancient economy is complex because the available evidence usually makes it impossible to verify hypotheses about the likely incentives resulting from various property rights regimes. Still, analysing ancient legal institutions can shed light on the basic relationships among the principal actors in an ancient economy, including the state, elite property owners, urban residents, and farmers.

Legal Institutions in the Greek World

One key issue is the role that legal institutions played in promoting commerce. The Greek world in the classical (480–323 BCE) and Hellenistic periods (323–31 BCE) was politically fragmented, and individual city-states (*poleis*) had their own legal systems. Consequently, we can speak of a unified system of Greek law only to a limited degree. This made it difficult to develop governance structures to enforce the types of contractual arrangements essential to commerce. In the Hellenistic period, the emergence of larger monarchies may have promoted a more unified system of commercial law between states. Eventually, the incorporation of the Greek world into the Roman Empire greatly enhanced the possibilities for developing a more uniform set of legal institutions. In the absence of unified formal institutions to govern commerce, we should expect merchants to have developed their own private ways of enforcing contractual obligations and resolving disputes (cf. Greif 2006).

At the level of the polis, we are best informed about the way in which commercial law functioned in classical Athens, particularly in the fourth century BCE (Todd 1993; Cohen 2005). At this time, Athens had become a commercial hub, and its involvement in commerce was vital to its survival, since it depended on imported grain from the Black Sea region. Certainly the economy of Athens, as much as any place in the ancient world, required legal mechanisms to develop and enforce complex commercial arrangements. The state intervened in commerce directly only to

protect the grain supply by imposing severe sanctions on Athenians who exported grain to other cities. Even so, institutions developed in Athens to promote trade. Banks played a crucial role in assembling the capital necessary for maritime commerce. Often these commercial undertakings might be complex, with multiple investors supplying cargoes to the same ship, so that a single voyage might involve a wide variety of contracts and loans (Cohen 1992). The question is how merchants involved in this commerce, many of whom came from locations overseas, could enforce the obligations of their trading partners. In most city-states, the local courts were open only to citizens, unless two states negotiated a bilateral commercial treaty. The Athenians endeavored to meet a more general need for a forum to resolve disputes by developing courts in which lawsuits involving overseas commerce, *dikai emporikai*, could be heard (Todd 1993, pp. 333–7; Cohen 2005, pp. 299–300). These courts were open to anyone doing business in Athens, not just Athenian citizens. Their success in encouraging commerce depended on their treating foreign traders in Athens impartially. Foreign traders, when sued in the court, had to post bond, but at the same time the courts discouraged frivolous lawsuits by imposing financial penalties on plaintiffs who failed to gain at least one-sixth of the jury's votes (Cohen 2005, pp. 299–302).

Another issue is the role that private contract law played in the economy. In contrast to Roman law, the mutual consensus of the two parties to a contract did not in and of itself create contractual obligations; rather, a real act, such as the exchange of property, was required (Todd 1993, pp. 262–8; Rupprecht 2005, p. 337). The apparent simplicity of this type of contract would seem to preclude certain complex commercial arrangements, such as sales of real estate on credit or sales of crops in advance of the harvest, but this was manifestly not the case. In Athens, one part of the solution to the problem was the freedom of procedure in courts; this flexibility made it possible to sue regardless of whether a business arrangement corresponded to an accepted contractual form.

The Hellenistic period saw legal developments potentially significant for the economy (Rupprecht 2005). Typically, multiple legal systems functioned side by side. In Egypt, for example, the Ptolemies, a Macedonian dynasty, introduced Greek law for the immigrant Greek population, while the native Egyptians continued to rely on their own legal traditions and contract forms. The substantial Jewish population in Egypt could also use its own laws. In Greek law, written documents were increasingly common in private business arrangements. They tended to be written in standard language, and so they would have served to make contracting in business simpler. The widespread use of written contracts means that there were scribes well versed in the basics of commercial law. The trend of using written documents to record what had originally been oral contracts accelerated in Roman times (Meyer 2004). A second development in Ptolemaic Egypt was the increasing registration of documents in state archives. In the early Roman imperial period in Egypt, the state developed a registry of real property and the rights assigned to it, the *bibliotheke enkteseon*, which helped to eliminate some of the uncertainty surrounding the ownership of real property that is characteristic of pre-modern economies.

The development of commerce in the Greek world, and in the Roman world later, depended on property owners having reliable agents to manage their businesses. Part of the solution in both the Greek and Roman worlds was to employ agents who were social dependants. In fourth-century Athens, this can be seen especially in the banking industry. The general prohibition against the ownership of land in Attica by non-Athenian citizens surely made banking an attractive business undertaking for resident aliens (metics), many of whom were quite wealthy. The foreign owners of these banks commonly employed slaves as their managers. A highly trained slave could operate a bank independently, but there was no threat that he would take advantage of his training to set up a rival bank to compete with that of his former employer (Cohen 1992, pp. 61–110, 133–6). In Rome, property owners employed slaves and freedmen in similar functions, as will be discussed below.

Legal Institutions in the Roman World

The development of Roman law as a legal system with wide application in the Mediterranean world had potentially enormous consequences for the Roman economy. Roman society had a professional class of jurists who interpreted the law in a rigorous fashion and, in effect, created a science of jurisprudence. The jurists originally provided legal advice in private trials, but beginning with the reign of Augustus (31 BCE – 14 CE) they gained a state-sanctioned role in providing authoritative interpretations of the law. In economic matters, one of the jurists' main contribution was to interpret contract law. By the second century BCE, the Roman praetors (the officials in charge of the administration of private law) had developed the concept of consensual contracts, including sale (*emptio-venditio*), lease and hire (*locatio-conductio*), mandate (*mandatum*), and partnership (*societas*). The contract types defined legal relationships crucial for the Roman economy, and they provided a basis for Roman commercial law for centuries to come. Although this is a controversial subject, it is now increasingly accepted that the jurists endeavored to respond to social needs as they interpreted contract law.

The Roman Empire was also successful in developing legal institutions that were accessible to a broad segment of the population. One key to this was the petition process. The Roman emperor received such a volume of petitions that the Roman government had an office, headed by an official of equestrian rank, the *a libellis*, whose responsibility was to receive petitions and issue answers, or rescripts, in the emperor's name (Peachin 1996). Petitioners would receive an authoritative response about the law applicable to their case, and they could then take these responses to local courts, whose judges would be obliged to follow them. People also sent petitions to officials of lower rank, from local magistrates to provincial governors. The petition process was so widespread that it suggests that the empire's subjects viewed it as a reasonably reliable way to protect their interests. Responding to petitions, moreover, provided the state with one way, albeit reactive, to intervene in the economy. Such intervention can be discerned in legal policies concerning farm tenancy, in such issues as the tenant's security of tenure and the allocation of the risks associated with agriculture (Kehoe 2007).

To consider agency again, the Romans, like the Greeks, often relied on social dependants, particularly slaves and freedmen, to serve as business agents. To some extent, this resulted from the basic organizing principles of the Roman household. In Roman society, the head of a Roman household, the *pater familias*, exercised a great deal of power over the members of his *familia*. These included his agnatic descendants as well as his slaves and freedmen. In economic terms, he was the ultimate owner of all the property in the hands of anyone in his power, or *patria potestas* (Saller 1994: 102–32). The *familia* provided the basic structure for organizing much of economic life in the Roman world. It was a setting in which people were trained in specialized skills important for the economy, and it also influenced the organization of commercial enterprises. When employing social dependants as agents, Roman property owners tended to give them a great deal of freedom. The slave would operate with a *peculium*, funds and property under his control but ultimately belonging to the owner. The slave agent had every incentive to manage the business well, since he could earn his freedom in doing so, whereas the owner could impose sanctions in the event of his misbehaviour more easily than would be possible with a free employee (Frier and Kehoe 2007, pp. 130–4). Often freedmen who gained their initial training as slaves could establish businesses of their own, training their own slaves, and continuing the cycle.

Merchants dealing with agents had to be assured that they would be able to enforce their claims in the event of a dispute. Part of the solution was a series of remedies, the so-called *actiones adiecticiae qualitatis*, created in the late third or second centuries BCE. These established the circumstances under which a property owner could be liable for obligations taken on by an agent. In many cases, the principal's liability was

limited to the size of the *peculium* granted the slave agent. This legal regime may have carried a substantial social cost, since in theory at least, the limited liability of the principal will have deterred some people from entering into otherwise productive business arrangements. At the same time, it responded to the needs of an upper class that was cautious in its approach to investing wealth (Kehoe 1997). The formal regime surrounding agency in Roman law can be contrasted with the type of agency that characterized Ptolemaic Egypt (Von Reden 2007, pp. 239–50). There, property owners who also held official posts relied on private, individual agents, who collected debts or made loans on their behalf. The activities of the agent, however, created no formal legal relationship between the property owner and a third party who was either a debtor or a creditor. This system of agency clearly revolved around the personal reputations of the individuals involved.

In interpreting Roman contract law, the Roman jurists seem to envision a class of independent contractors who had sufficient resources to undertake major jobs, such as leasing farms or construction projects. In the contractual relationship covering major construction projects, called *locatio-conductio operis* (Martin 1989), the builders were expected to organize tasks and finance operations until they were paid by their principals. Again, this situation can be contrasted usefully with the corresponding contract arrangement in Ptolemaic Egypt, called *ergolabia*. In such contracts from the third century BCE, for example, the property owner employing the contractor generally had to pay the latter up front. The contractor still had a great deal of responsibility, but the payment up front created potential monitoring problems, and it was probably necessary because at this time contractors did not have ready access to cash (Von Reden 2007, pp. 146–50).

This brief survey suggests some of the issues that can be investigated by a careful analysis of the relationship between legal institutions and the economy in the ancient world. By investigating legal institutions, we can better understand better the relationships that shaped the economy and the likely implications of these relationships for economic performance.

See Also

- ▶ Agency Problems
- ▶ Ancient Greece, The Economy of
- ▶ New Institutional Economics

Bibliography

Cohen, E.E. 1992. *Athenian economy and society: A banking perspective*. Princeton: Princeton University Press.

Cohen, E.E. 2005. Commercial law. In Gagarin and Cohened. 290–302.

Frier, B.W. and D.P. Kehoe. 2007. Law and economic institutions. In Scheidel et al. ed., 113–143.

Gagarin, M., and D. Cohen, ed. 2005. *The cambridge companion to ancient greek law*. Cambridge: Cambridge University Press.

Greif, A. 2006. *Institutions and the path to the modern economy: Lessons from medieval trade*. Cambridge: Cambridge University Press.

Kehoe, D.P. 1997. *Investment, profit, and tenancy: The jurists and the Roman agrarian economy*. Ann Arbor: University of Michigan Press.

Kehoe, D.P. 2007. *Law and the rural economy in the Roman empire*. Ann Arbor: University of Michigan Press.

Martin, S.D. 1989. *The Roman jurists and the organization of private building in the late republic and early Empire*. Brussels: Collection Latomus 204.

Meyer, E.A. 2004. *Legitimacy and law in the Roman World: Tabulae in Roman belief and practice*. Cambridge: Cambridge University Press.

Peachin, M. 1996. *Iudex vice caesaris: Deputy Emperors and the administration of justice during the principate*, Heidelberger Althistorische und Epigraphische Beiträge. Vol. 21. Stuttgart: Steiner.

Rupprecht, H.-A. 2005. Greek law in foreign surroundings: Continuity and development. In Gagarin and Cohen ed. 328–342.

Saller, R.P. 1994. *Patriarchy, property and death in the Roman family*. Cambridge: Cambridge University Press.

Scheidel, W., I. Morris, and R. Saller, ed. 2007. *The Cambridge economic history of the Greco-Roman World*. Cambridge: Cambridge University Press.

Todd, S.C. 1993. *The shape of athenian law*. Oxford: Clarendon Press.

Von Reden, S. 2007. *Money in ptolemaic egypt: From the Macedonian conquest to the end of the third century BC*. Cambridge: Cambridge University Press.

Lehfeldt, Robert Alfred (1868–1927)

S. Herbert Frankel

Lehfeldt was born in Birmingham on 7 May 1868 and died in Johannesburg on 11 September 1927. He obtained the BSc degree from London University in 1889 and a BA in 1890 from Cambridge, where he was at St John's College. In 1906 he accepted the Chair of Physics in the South African School of Mines and Technology, Johannesburg. In 1913 he exchanged this for the new chair of economics at the University of the Witwatersrand. Lehfeldt was undoubtedly the leading economist in South Africa and was the correspondent of the Royal Economic society for South Africa. He contributed frequently to the *Economic Journal* and to the *Journal of the Royal Statistical Society.*

Lehfeldt was a brilliant mathematician, statistician and demographer, and a pioneer in the application of mathematical analysis to economic and social problems. He was an international authority on currency questions, on the economics of gold mining, on the relation of the world's supply of gold to the course of prices, and on the monetary role of gold. Regulating and stabilizing long-period changes in the value of gold, by international control of the supply rather than of the demand for the metal, was advocated in his *Restoration of the World's Currencies* (1923).

In South Africa he gave evidence to many Commissions of Inquiry. He was largely responsible for the creation of the Economic Society of South Africa and was also a member of the Statistical Council. In *The National Resources of South Africa* (1922) he was the first to estimate the national income of the country and to assess the contribution of the Coloured and Native population to the economy.

Selected Works

1919. *Gold prices and the witwatersrand*. London: P.S. King & Son.
1922. *The national resources of South Africa*. Johannesburg/London: University of the Witwatersrand Press/Longmans, Green & Co.
1923a. Return to capital invested in the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*.
1923b. *Restoration of the world's currencies*. London: P.S. King & Son.
1926a. *Money.* London: Oxford University Press.
1926b. *Controlling the output of gold*. With a Preface by Dr Hjalmar Schacht. London: General Press.
1927. *Descriptive economics*. London: Oxford University Press.

Lehman Brothers Bankruptcy, What Lessons can be Drawn?

Thomas J. Fitzpatrick IV and James B. Thomson

Abstract

There is disagreement about whether large and complex financial institutions should be allowed to use US bankruptcy law to reorganise when they get into financial difficulty. We look at the events surrounding the Lehman Brothers bankruptcy filing for lessons as to whether bankruptcy law could be used to produce an orderly windup of the affairs of a failed financial firm. If so, then judicial resolution under the US Bankruptcy Code might be a better alternative to bailouts or to resolution under the Dodd–Frank Act's orderly liquidation authority. We find that there is no clear evidence that bankruptcy law is insufficient to handle the resolution of large complex financial firms.

Keywords

Bankruptcy; Causation; Contagion; Dodd–Frank; Insolvency; Orderly resolution authority

JEL Classifications

E44; G01; O11; K20; K22; G28; G38; G18; G21; G28; N12

What is in a name? In the case of Lehman Brothers the name has two different and distinct meanings. Prior to the autumn of 2008, Lehman Brothers referred to one of the oldest investment banks in the USA, with roots in the cotton exchange of the mid-19th century. At the time it filed for protection under Chapter 11 of the US Bankruptcy Code, Lehman Brothers Holdings International was the fourth largest US investment bank and the largest bankruptcy on record. Today Lehman Brothers, used synonymously with the Lehman Brothers bankruptcy filing, is commonly used to refer to an important episode during the 2007–2009 financial crisis. To borrow a line from Winston Churchill, the Lehman Brothers bankruptcy filing on 15 September 2008 did not represent the beginning of the end of the financial crisis, but rather marked the end of the beginning.

Just the Facts

In the 1960s police drama *Dragnet*, the main character Sergeant Joe Friday would direct witnesses to give him 'just the facts'. So what are the facts concerning the episode of the financial crisis attributed to the Lehman bankruptcy?

The Lehman Brothers bankruptcy filing occurred during a period of market turmoil which intensified in the days that followed. Financial markets continued to exhibit signs of increased stress thereafter and during the autumn of 2008. Yields in short-term markets spiked during the week following the Lehman filing. Risk spreads in short-term credit markets widened –indicating a 'flight to quality' by market participants. For example, the 3-month term LIBOR-OIS spread, an indicator of market stress (Thornton 2009), increased around 14.75 basis points from the Friday before the Lehman bankruptcy filing to 16 September, the day after. From 16 September to 10 October the LIBOR widened by another through 263 basis points. Increased market stress was also evident in the credit default swaps (CDS) market, where the cost of buying credit protection rose sharply in the days just after the Lehman Brothers bankruptcy filing. The five-year CDX.NA.IG index (which is an index of credit default swaps written against North American investment grade companies from Markit and Bloomberg) rose 55 basis points, a 36% increase from 12 September to 17 September. The CDX.NA.IG index declined from its 17 September peak to the end of the month, but still finished September some 20 basis points higher than where it started.

The financial turbulence in the autumn of 2008 was the product of a series of events. The Lehman bankruptcy was one of nearly two dozen significant disruptive events in September 2008 alone, some unrelated to the Lehman bankruptcy filing and some related to its failure. Notable among the economically significant events is the placement of Fannie Mae and Freddie Mac in conservatorship by the Federal Housing Finance Authority, the Federal Reserve assisted rescue of AIG by the US Treasury, and the deathbed acquisition of Merrill Lynch by Bank of America Corporation. Also, notable is the Reserve Primary Money Fund announcement that it had 'broken the buck': due to losses on its holdings of Lehman debt, the net asset value of the Fund's shares had fallen to $0.97 a share. It was only the second time since the SEC adopted rules governing money market mutual funds in 1983 that a money market fund's share value had fallen below one dollar. Runs on money market mutual funds (MMMFs) would follow.

Interpreting the Facts

While the facts about what happened and when are clear, the connections between them are not. Drawing inferences from any single event is problematic at best. Just as any single point on a plane is consistent with an infinite number of lines, a single event may not allow one to discriminate between numerous different hypotheses. Not surprisingly, there are two different interpretations of the facts associated with Lehman and they arrive at diametrically opposed positions as to causation, and the implications of it for the use of the Bankruptcy Code to handle failing financial firms.

One of the most contentious issues emanating from the Lehman Brothers episode is whether the bankruptcy process is, or with modifications

could be, a suitable method for handling the failure of complex, non-bank financial firms. Opinions are sharply divided on the adequacy of US bankruptcy law to resolve complex non-bank financial firms in an orderly fashion. Bankruptcy scholars argue that the market turmoil in the aftermath of the Lehman bankruptcy had little to do with the use of bankruptcy to resolve it, and that in the face of the complexity inherent in resolving an institution the size and scope of Lehman Brothers, the bankruptcy was orderly. In other words, there was no causation running from the bankruptcy filing to the disorderly markets that followed. Proponents of this view argue that the near collapse of markets following Lehman's bankruptcy filing was the result of policy uncertainty: The US government decided to let Lehman fail when the market expected a government-assisted rescue. In fact, Lehman was not prepared for its bankruptcy filing, ostensibly because its management expected government intervention to prevent this outcome (Miller 2010).

The other view, which one might call the official view of the Lehman episode, is that Lehman's filing for protection is articulated by the Federal Deposit Insurance Corporation (FDIC), among others, which interpreted the facts as supporting a causal relationship between the financial turmoil following Lehman's bankruptcy filing and the use of bankruptcy to resolve Lehman. Under this view, the near collapse of markets in the days following the bankruptcy filing was a direct result of a disorderly windup of Lehman's affairs. Under this interpretation of events in the autumn of 2008 the answer is clear – an orderly resolution of the insolvency of a large financial firm cannot be done in bankruptcy.

This debate is largely unsettled. Even the Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010 (DFA) appears to codify both positions. Title II of DFA creates the Orderly Liquidation Authority (OLA), an administrative receivership process under the FDIC to resolve systemic financial companies. OLA is, however, an exceptional power for resolving systemic non-bank financial firms; bankruptcy remains the default. In addition, DFA mandates that systemic financial companies create and maintain 'living wills': resolution plans for dismantling them in bankruptcy.

Understanding the lessons of the episode during the financial crisis identified with the Lehman Brothers bankruptcy filing requires a careful accounting of the cluster of events that surrounded it. Moreover, no analysis would be complete without an analysis of the role of incentives and expectations in the setup and propagation of the financial crisis. Studying the entire mosaic of the Lehman Brothers episode is necessary to provide context to the period in question and proper attribution of the effects of the bankruptcy filing on the subsequent market turmoil.

As the Lehman episode represents one point in financial history it is impossible to prove or disprove any reasonable interpretation of it. It is possible to, however, to point to some lessons that can be drawn from it. These lessons concern whether the insolvency of large or complex financial companies can be adequately handled through the judicial process of bankruptcy. Moreover, an understanding the Lehman Brothers episode may point to types of reforms to the Code that may be required if bankruptcy is to be a viable option for handling large complex financial firms and a desirable alternative to *ad hoc* bailouts or to resolution under the DFA's Orderly Liquidation Authority.

International Issues

Every country's insolvency regime is inherently complicated by its jurisdictional boundaries. Systemically important financial institutions do not operate in a single country, nor do they have all of their assets located in a single jurisdiction. When Lehman filed for bankruptcy, it operated nearly 3,000 US and foreign chartered separate entities in 20 countries, and its complex legal structure was virtually unrelated to its operational structure (Cumming and Eisenbeis 2010). This made it incredibly difficult to determine what assets were in each entity in a bankruptcy estate. Further complicating this, substantial sums were transferred between Lehman's cross-border subsidiaries on the eve of bankruptcy.

While Lehman's global presence added substantial complexity to the resolution process, it is difficult to argue that this complexity is a shortcoming of US bankruptcy law. US bankruptcy law has provisions to address cross-border insolvencies (Chapter 15), but these do not guarantee effective or efficient operation. Each country has its own insolvency regimes, and there is substantial variation in their treatment of creditors. This is an issue present whenever a global institution is resolved under any bankruptcy scheme, and to date very little has been done to address it. The United Nations Commission on International Trade Law has developed a model law on cross-border insolvency, but it has not yet been adopted by a sufficient number of jurisdictions to be meaningfully operable. In some sense the international issues raised by the failure of Lehman is immaterial to the insolvency regime debate in the USA. Nonetheless, one lesson that can be learned from the Lehman bankruptcy is that there is plenty of room for improvement in cross-border insolvency regimes.

US Bankruptcy Law and Complex Financial Institutions

Irrespective of international issues, some analysts maintain that it was Lehman's use of the bankruptcy courts that caused the market turmoil. They often point to the increased financial turmoil during the week following Lehman's bankruptcy filing as evidence of the insufficiency of bankruptcy law to resolve complex financial firms. Others claim that it was not the use of bankruptcy, but rather policy responses inconsistent with market expectations that caused markets to panic. That is, Lehman was allowed to fail when financial markets, and even the Lehman management team, expected a government-assisted rescue. A closer look at events around that time suggests that neither view is entirely correct.

The Lehman bankruptcy occurred during a time when there were good reasons for market participants to question the solvency of a number of large financial firms. As noted above, the bankruptcy was accompanied by nearly two dozen significant disruptive events in September 2008 alone. The clustering of multiple events around the time of the bankruptcy makes it difficult to identify the causal effects of the bankruptcy on markets, let alone the effect of the use of US bankruptcy law.

While Lehman's failure triggered many problems in markets, event clustering makes it impossible to identify empirically the use of bankruptcy courts as the root of those problems. Moreover, it is impossible to separate out the impact of Lehman's bankruptcy filing from the uncertainty created by its filing.

Studies have shown that such uncertainty can have significant effects on markets. For example, in 1982 Penn Square Bank was liquidated by the FDIC, which experimented with modified payouts to resolve large bank failures (Furlong 1984). These modified payouts created uncertainty in the minds of the large, explicitly uninsured creditors of Continental Illinois as to whether they were exposed to losses in the event Continental was closed. This uncertainty drove the run on Continental Illinois' deposits before its collapse in 1984 (Sprague 1986).

The source of market turmoil following Lehman's failure, then, cannot conclusively be attributed either to the use of bankruptcy law to resolve the firm's insolvency or to the uncertainty created by policy actions inconsistent with market expectations.

Bankruptcy and Contagion

When a large, complex financial firm fails, the method of resolution should not be conducive to contagion. That is, the resolution process should not endanger the solvency of other firms. This is especially true in systemic crises, when the financial system is already stressed. Bankruptcy critics often argue that bankruptcy law may trigger contagion because it is designed to pay creditors strictly according to the priority of their claims. There is no consideration of their financial condition or potential market instability. Thus, contagion may spread through the use of bankruptcy if the recovery of creditors in need of liquidity is insufficient, or indirectly through CDS written on

the resolved firm's debt. But the Lehman bankruptcy does not support the view that bankruptcy leads to contagion.

As mentioned above, the day after Lehman Brothers filed for bankruptcy, the Reserve Primary Money Fund announced that it had 'broken the buck': this reflected how large an impact Lehman's collapse was having.

Most analysts would concede that the Fund's 'breaking the buck' was a direct consequence of the Fund's losses on its holdings of Lehman debt, that the losses led to contagion, and that the contagion effects impacted the money market mutual fund industry and the commercial paper market thereafter. It is harder to argue that the structure of US bankruptcy law, and not the insolvency of Lehman itself, was responsible for the losses on Lehman debt and the subsequent contagion. It may also be the case that the contagion effects were more a consequence of the money market funds' overexposure to Lehman and to a specific feature of the money funds themselves – the pegging of the share price to $1. The share-price peg creates incentives for retail customers to run on a fund when its ability to maintain the peg becomes uncertain. Customers believe it is in their best interest to run to ensure par redemption of their money-fund shares.

Lehman's bankruptcy also tested the CDS market, as there was a reported $400 billion of credit protection written against Lehman's debt. At the time of its bankruptcy, Lehman was the largest failure to be handled in the CDS market. For the purpose of settling the CDS contracts, Lehman's debt was determined to be worth 9.75 cents on the dollar at an International Swaps and Derivatives Association auction, lower than the pre-auction estimates of 12 to 15 cents. However, the settlement of credit protection written on Lehman did not have material effects on financial markets (Summe 2009; Senior Supervisors Group 2009).

Bankruptcy and Qualified Financial Contracts

Derivatives and repos are special types of contract called qualified financial contracts (QFCs), which are exempt from the trust avoidance powers of the Bankruptcy Code and the automatic stay. The trust avoidance provisions and automatic stay are designed to coordinate creditor payouts and ensure that they occur according to the priority of the claims that existed when the original agreements were made. These provisions are designed to prevent a race to grab a firm's assets on the eve of failure or after the firm fails. Instead of being stayed and handled through the bankruptcy estate, each counterparty may close out, net, and settle its QFCs before other debts are paid in bankruptcy. In a sense, QFCs are super priority claims, as they are settled before all others. The special treatment of QFCs may complicate the process of reorganising financial companies in bankruptcy by allowing counterparties to grab assets before the claim priority provisions take hold, but bankruptcy experts disagree about the effect of the QFC exemption in bankruptcy. There is even disagreement on how well Lehman's QFC book, the largest in history to be handled in bankruptcy, was dealt with.

While Lehman's reorganisation has provided additional guidance on which financial contracts are exempted from the automatic stay and how QFCs will be handled in bankruptcy, there is still disagreement on how well bankruptcy handles QFCs. Generally opinions fall into one of two schools of thought. First, there are those who argue that the QFC exemption was an obstacle to an orderly resolution in the Lehman case. In testimony before a House subcommittee in 2009, Harvey Miller, the lead bankruptcy attorney for Lehman, argued that the exemption of some 930,000 derivative counterparties from the automatic stay led to a massive destruction of value through counterparties canceling their contracts. Ayotte and Skeel (2010) and Roe (2011) argue that the safe harbour provisions of bankruptcy for QFCs create perverse incentives for counterparties. Those incentives contribute to the systemic implications of a firm's failure, including creating a stampede for the exits, which inhibit orderly resolution under bankruptcy.

Second, there are those who argue that Lehman's derivatives portfolio was handled effectively *because of* the exemption from the automatic stay. Kimberly Anne Summe, a former managing director at Lehman, provided this

interpretation of the impact of Lehman's counterparties cancelling their contracts on the value of Lehman's estate. Summe noted that only around 3% of Lehman's derivative contracts remained in the bankruptcy estate 106 days after the filing, potentially preventing the spread of distress to Lehman's counterparties by allowing them to close out quickly and re-establish their hedges before market conditions changed too dramatically (Summe 2009). However, the benefit of allowing quick re-hedging is unclear, as is the cost of losing going-concern value (the value of the company as an ongoing entity rather than a liquidated one) due to the stay exemption.

To the extent that the Bankruptcy Code's safe harbour provisions for QFCs are a stumbling block to an orderly resolution of a systemic financial firm, a simple amendment to the Code is the logical fix. In fact, bankruptcy supporters argue for such a change in the law subjecting QFCs to a limited automatic stay, and there appears to be a case for their position. The FDIC enjoys a one-day stay on QFCs in bank receivership cases, and there is little evidence that this limited stay for FDIC receiverships has been a problem. Moreover, when a non-bank financial firm is resolved under the orderly liquidation authority established in the Dodd–Frank Act, QFCs are subject to a one-day stay. Both provisions allow for the transfer of QFCs during the stay. If this stay is priced into QFCs with depository or systemically important financial institutions and US bankruptcy law were changed to parallel the Dodd–Frank provision, markets would not likely be disrupted, and the pricing of QFCs would be identical across counterparties. It would also have the added benefit of giving the bankruptcy estate up to three days to determine what to do with a derivatives book before counterparties could close out and net, provided that the insolvent firm filed on a Friday.

The Scope of US Bankruptcy Law

The final material stumbling block to an orderly resolution under bankruptcy of a complex financial firm such as Lehman is the exclusion of certain types of businesses from Chapter 11 (which provides for corporate reorganisation). In the case of Lehman, the exclusion of its broker-dealer subsidiary (Lehman Brothers, Inc.) from filing for Chapter 11 complicated the resolution of Lehman Brothers Holdings International. Lehman Brothers, Inc., became the subject of a liquidation proceeding under the US Securities Investor Protection Act four days after Lehman Brothers Holdings International filed for bankruptcy, during which time the brokerage was borrowing from the Federal Reserve Bank of New York under the Primary Dealer Credit Facility.

The absence of government support likely would have complicated the sale. Because it did not have access to the special financing provisions that firms filing under Chapter 11 are entitled to, the brokerage would have lost going-concern value but for its access to the Primary Dealer Credit Facility. While the sale of Lehman's broker-dealer to Barclay's was quickly approved, without government support the sale might not have been possible under bankruptcy law. Whether this merits a change in US bankruptcy law would have to be addressed separately for each exemption, though some argue that the prohibition of broker-dealers reorganising in bankruptcy no longer makes sense (Skeel 2010).

Policy Implications

Lehman Brothers Holdings International is not the first, nor likely the last, systemic financial company to run aground. The case is interesting, however, because the failure occurred during the most severe financial crisis in the USA since the Great Depression. The economic and financial market climate in which Lehman failed greatly complicated any resolution method that did not involve taxpayer assistance in the form of capital infusions or blanket guarantees of creditors. Yet Lehman became the poster child for the orderly liquidation authority provisions of Title II of the 2010 Dodd–Frank Act.

Drawing inferences from Lehman about the effectiveness of bankruptcy in dealing with failing financial firms is problematic. It is difficult to use a single data point – the Lehman bankruptcy – to separate out the impact of Lehman's failure, the

use of bankruptcy to resolve it, and the policy uncertainty.

Still, Lehman's bankruptcy offers guidance on how to approach future failures of large, complex financial firms. It appears that there are provisions of bankruptcy law that merit review and possible revision. In the absence of those changes, it may be the case that systemically important pieces of an insolvent firm may be more effectively resolved in an administrative proceeding such as the Orderly Liquidation Authority established under Dodd–Frank. But based on the experience with Lehman, there is no clear evidence that bankruptcy law is insufficient to handle the resolution of large, complex financial firms.

See Also

- ▶ Bankruptcy Law, Economics of Corporate and Personal
- ▶ Bankruptcy, Economics of
- ▶ Credit Crunch Chronology: April 2007–September 2009
- ▶ Deposit Insurance
- ▶ Fall of AIG
- ▶ Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance
- ▶ Federal Reserve System
- ▶ Finance (New Developments)
- ▶ Financial Market Contagion
- ▶ Minsky Crisis
- ▶ Regulatory Responses to the Financial Crisis: An Interim Assessment

Bibliography

Ayotte, K., and D.A. Skeel Jr. 2010. Bankruptcy or bailouts? *The Journal of Corporation Law* 35(3): 469–498.

Board of Governors of the Federal Reserve System. 2011. *Study on the resolution of financial companies under the Bankruptcy Code*.

Cumming, C., and R.A. Eisenbeis. 2010. *Resolving troubled systemically important cross-border financial institutions: Is a new corporate organizational form required? Federal Reserve Bank of New York Staff Report*, vol. 457.

Edwards, F., and E.R. Morrison. 2005. Derivatives and the bankruptcy code: Why the special treatment? *Yale Journal on Regulation* 22: 101–132.

Federal Deposit Insurance Corporation (FDIC). 2011. The orderly liquidation of Lehman Brothers Holdings Inc. Under the Dodd–Frank Act. *FDIC Quarterly* 5(2): 1–19.

Furlong, F.T. 1984. FDIC's modified payout plan. *Federal Reserve Bank of San Francisco Economic Letter*, 18 May.

Mengle, D.L. 2007. Credit derivatives: An overview. *Federal Reserve Bank of Atlanta Economic Review* 92(4): 1–24.

Miller, H. R. 2009. Testimony before the Subcommittee on Commercial and Administrative Law of the House of Representatives Committee on the Judiciary, 111th Congress, 1st Session, for Hearings on 'Too big to fail: The role for bankruptcy and antitrust law in financial regulation reform', 22 October.

Miller, H.R. 2010. Testimony before the financial crisis inquiry commission for hearing on examining the causes of the current financial and economic crisis of the United States and of the collapse of Lehman Brothers, 22 October.

Roe, M.J. 2011. The derivatives players' payment priorities as financial crisis accelerator. *Stanford Law Review* 63: 539–588.

Senior Supervisors Group. 2009. *Observations on management of recent credit default swap credit events*, Available at: http://www.sec.gov/news/press/2009/report030909.pdf.

Skeel, D.A. 2010. Bankruptcy boundary games. *Brooklyn Journal of Corporate, Financial and Commercial Law* 4: 1–21.

Skeel, D. 2011. *The new financial deal: Understanding the Dodd–Frank Act and its (Unintended) consequences*. Hoboken: John Wiley & Sons.

Sprague, I.H. 1986. *Bailout: An insider's account of bank failures and rescues*. New York: Basic Books.

Summe, K.A. 2009. Chapter 5, Lessons learned from the Lehman bankruptcy. In *Ending government bailouts as we know them*, ed. K.E. Scott, G.P. Schultz, and J.B. Taylor. Stanford: Stanford University Press.

Thornton, D.L. 2009. What the Libor-OIS spread says. *Federal Reserve Bank of St. Louis Economic Synopses*,vol. 24.

Leisure

Lars Osberg

Abstract

Economists have typically defined 'leisure' residually, as equal to 'non-work time', and, despite the problematic classification of enjoyable jobs, commuting time and unemployment,

presumed that individuals derive utility from non-work time and disutility from working time. However, a recent literature now emphasizes 'social leisure' and coordination problems in leisure time. Since longer working hours by some individuals make arranging a social life more difficult for others (thereby decreasing the utility of their non-work time), externalities in time use may create multiple possible equilibria in time use, which may explain the sharp divergence in working hours between Europe and the United States.

Keywords

Becker, G.; External economies; Goods-intensive commodities; Individualism; Labour supply; Leisure; Marriage; Social capital; Social interaction; Time use; Time-intensive commodities; Unemployment; Well-being

JEL Classifications

D11

What is 'leisure'? The Merriam-Webster Online Dictionary defines it as 'freedom provided by the cessation of activities; *especially*: time free from work or duties', while the Oxford English Dictionary suggests it is 'The state of having time at one's own disposal; time which one can spend as one pleases; free or unoccupied time'. (Both note that the adjective 'leisurely' describes an action that is done without haste, in a relaxed way.) In common parlance, attendance at a relative's funeral or time spent voting would therefore not generally be seen as 'leisure', because time spent on an activity due to a sense of civic or familial duty cannot qualify.

'Leisure' is therefore a problematic concept for economists, because the context and subjective interpretation of an activity is crucial to deciding whether it should be counted as work, duty or leisure – cooking or driving are, for example, activities that may be performed as parts of a paid occupational role, as a duty or for personal enjoyment. It is, in fact, not easy to think of an activity or time use that is not done sometimes for pay, sometimes for duty and sometimes for pleasure – perhaps by different people, but sometimes also by the same people. In many universities, the subtleties of such distinctions are explored in departments of 'Leisure Studies', which is now a recognized area of academic teaching and research. Peer-reviewed journals such as *Annals of Leisure Research* or *Leisure Sciences* report the latest research on leisure activities, and conferences are organized on such topics as 'Serious and Casual Leisure'.

Leisure as a Residual Category: The Standard Approach

However, for many economists, 'leisure' is simply the L in labour supply theory. This approach starts, in a one-period model, with each individual maximizing a utility function, where U is the individual's utility level, C represents consumption goods and L is leisure time, as in Eq. 1:

$$\text{Max } U = u(C, L)[u' > 0 u'' < 0] \tag{1}$$

The wage rate available in the paid labour market (w) and total time (T) are seen as the fundamental constraints facing individuals. In this framework, the problem of utility maximization can be equivalently seen as one of 'labour supply' or 'leisure demand' since total time is divided between hours of paid work (H) and leisure time (L).

$$H + L = T \tag{2}$$

$$C \leq wH. \tag{3}$$

From this perspective, 'leisure' is whatever 'work' isn't – that is, leisure is a residual category, which is rarely examined directly or defined explicitly. Standard practice in economics journals is to focus on the hours of work decision – and 'work' is usually interpreted to mean 'paid employment'. In the JSTOR database of the top 26 economics journals, a keyword search, conducted in July 2005, for 'leisure' in archived articles published since 1995 yielded 823 'hits'. Of the top 100, sorted for 'relevance', only 25 had an explicit verbal definition of leisure – in most

cases leisure was defined implicitly, as in Eq. 2. If one discards the three articles discussing consumer demand for 'leisure goods' and focuses on time use, one finds the overwhelming majority of articles used leisure as a synonym for 'non-market time' – only three per cent recognized the possibility of 'on the job leisure' (but the definition was similarly residual – a lack of work effort – and implicit – for example, Dickinson 1999, p. 639). Relatively few articles (about 15 per cent) considered the possibility that home production (such as shopping time) may be a form of 'work', while a similar number (about 13 per cent) argued that time spent in schooling or training preparatory to paid employment is not leisure. For a very few articles (three per cent), leisure was the residual time available after paid work and some other alternative, such as criminal activity.

When working time is defined as equal to hours of paid employment, commuting time is implicitly defined as part of leisure, although it is plausibly an intermediate input into paid employment. Commuting time is an important percentage of time use in modern societies – Putnam (2000, p. 212), for example, has ascribed much of the decline in civic engagement in the United States to increased commuting time and commented that 'American adults average seventy-two minutes every day behind the wheel....more than we spend cooking or eating and more than twice as much as the average parent spends with the kids'. However, commuting time is strangely absent from most labour–leisure models. As well, although 'retirement' is the particular form of non-work time consumed at the end of the life cycle, most economics articles implicitly exclude it from analysis, by concentrating on the working-age population.

All the same, although $L = T - H$ remains the dominant approach in economics, it has long been recognized that classifying time use as 'work' (painful) or 'leisure' (pleasurable) can be a bit oversimplified. A large body of research indicates, for example, that the unemployed are typically quite unhappy (Frey and Stutzer 2002; Di Tella et al. 2003) – time spent in unemployment seems to be qualitatively different from non-work time spent in other ways (that is, unpleasant). In general, people tend to rank their jobs fairly highly when asked to compare the satisfaction derived from specific activities (including jobs and types of housework and leisure). Juster and Stafford (1985) argued long ago that, in general, activities that involve social interaction – whether paid or unpaid –tend to be highly valued by individuals. Gary Becker (1965, p. 504) commented even earlier that 'Not only is it difficult to distinguish leisure from other non-work, but also even work from non-work'.

'Time-Intensive Commodities' and the Disappearance of 'Leisure': The Becker Approach

Becker's solution to the time classification problem was to posit that 'commodities' (like dinner, or a sailing excursion) are what enters individuals' utility functions, and that the production of these commodities requires the input of both material goods and time. In this approach, 'leisure' therefore disappears as a distinct category, somewhat replaced by the concept of a 'time-intensive commodity'. The Becker perspective has important implications for the type of leisure activities that people are predicted to choose. Personal time is, essentially, the only input into commodities like contemplation or conversation or the pure enjoyment of peace and quiet – so their cost is just the opportunity cost of time (that is, the wage rate). The cost of goods-intensive non-work commodities (like speedboat racing) depends partly on the cost of those material goods. When (if) the wage rate rises, time-intensive leisure commodities increase in relative price compared with goods-intensive commodities. Hence, the Becker prediction is for greater materialism over time.

As well, consuming more 'commodities' in the same time period – for example, squeezing a tennis game and a sail and dinner and a night at the opera into the same day – is seen in the Becker model as representing an increase in the 'productivity of consumption time' (and more is always better), but some would also describe this as a more frenetic lifestyle. Winston (1987, p. 160) has commented that 'the most serious casualty

[in Becker's approach] was loss of the sense of a leisurely and controlled pace that produces genuine satisfaction'.

However, Becker's approach has not, in fact, been much used. The straightforward work–leisure dichotomy continues to dominate economics journals. The pleasures of non-work time and the marginal disutility of labour were stressed by Marshall (1920, p. 117) many decades ago, and they continue to be the dominant framework today. Can one – should one – expect this constancy of perspective among economists to persist?

Social Leisure and the Coordination Problem

One of the peculiarities of the traditional 'leisure demand–labour supply' perspective is its individualism. If utility really did depend only on the quantity of consumption goods and number of non-work hours experienced by individuals, a person's level of utility would be unaffected by solitary confinement, or by any other configuration of social interaction. However, time spent in isolation is, for most people, pleasurable only in small doses. Although one can choose to be alone, relatively few leisure activities are intrinsically asocial. Most leisure activities can be arranged on a continuum of 'teamness', and the vast majority of them are distinctly more pleasurable if done with others.

Playing softball or soccer are activities that make no sense if done alone. Singing to oneself may be something done in the shower, but singing with a choir is generally a different level of experience. Travelling to exotic foreign places or going for a walk are activities which are usually more pleasurable if done with a companion. Reading a novel is certainly solitary, but many people also like to talk about it afterwards, either formally in a book club or informally with friends over dinner.

To list these different possible leisure activities is to underscore the variety of leisure tastes that individuals have. This variety creates, for each individual, the problem of locating somebody congenial to play with, and scheduling the simultaneous free time to do so. The basic problem with wanting to have a social life is that individuals cannot do it unilaterally – arranging a social life involves a search process which is constrained by the social contacts available to each person, and by the availability of other people. This interdependence of leisure has generated a new literature, with a set of new insights.

Corneo (2005), for example, contrasts privately consumed leisure time (watching television) and socially enjoyed leisure (which requires investment in relationships). Across nations, average hours of television watching are *positively* correlated with average working time. Corneo explains this in terms of the strategic complementarities that arise in the organization of social leisure. If these complementarities are strong enough, equilibria with little social leisure but long hours of work and television viewing, and equilibria in which there is much social leisure along with short hours of work and television viewing, are both possible. Although workers will prefer the higher wages and lower hours of work of the latter, capitalists will prefer the former, since they realize a higher rate of return on their capital stock when total hours of work increase. And if desired working hours are conditional on what others do, individuals need coordination devices to ensure that social leisure is feasible – such as public holidays, a common weekend or working hours regulation – which implies a potentially crucial role for the state and for the relative power of workers and capitalists in influencing public policy.

Jenkins and Osberg (2005) argue that, although solo television watching is certainly feasible, companionship may nonetheless increase the utility derived from the activity. Their emphasis is on modelling more explicitly the constraints involved in locating leisure companions. They argue that the leisure time choices of household members depend on the opportunities for associational life that exist outside the household, and they show that the likelihood of associational activity for persons of a given age group depends on the percentage of persons in other age groups that also engage in that activity. They note that economic models of marriage have discussed the

interdependence of spouses in income and material consumption, but it is also plausible that an important reason for marriage is that couples may like spending time together. Like Hamermesh (1998, 2002), they provide evidence on the synchronization and scheduling of spousal work and leisure time.

What are the implications of these new models of social leisure? From a theoretical perspective, the emphasis on the social nature of leisure opens up a whole new set of coordination issues – there is certainly no presumption that individualistic decision making will automatically produce a socially optimal equilibrium. However, the new models of social leisure nest the old labour–leisure choice perspective, since the option of 'solo leisure' is always there (albeit now one of several alternatives).

Kuhn (1970) argued that paradigms are replaced when they confront an important empirical anomaly that they are unable to resolve and when a more encompassing alternative theoretical perspective becomes available. The empirical fact which is now forcing a reconsideration of the analysis of leisure is the huge size of cross-national differences in the trend and level of non-work time. From 1980 to 2000, for example, average annual working hours per adult (ages 15–64) rose by 234 hours in the United States to 1,476 hours, but fell by 170 hours in Germany to 973, and by 210 hours in France to 957 (see Osberg 2003a). By 2000, the cross-sectional difference was huge – non-work time per adult per week was some 9.7 hours greater in Germany, and 9.9 more hours greater in France, than in the United States.

In principle, an increase in hourly wages increases both potential income and the opportunity cost of leisure, so the demand for a normal good (like leisure) may rise or fall depending on the relative size of income and substitution effects. However, why should one be larger in Europe and the other larger in America? It is just not very satisfactory to say that 'tastes differ'.

Cross-country differences in average leisure time are due in part to inter-country differences in probability of employment, in part to differences in common entitlements to paid vacations and public holidays, and in part to differences in the usual hours of work of employees. Trends in these three components are driven by distinctly different processes – the number of paid public holidays is, for example, determined by a set of political processes quite different from the determinants of individual decisions to enter the workforce and to work specific hours. A robust debate has emerged over the causes of these differences in total leisure time (for example, Bell and Freeman 2001; Alesina et al. 2005) – but it is clear that these differences are large enough to motivate both a concern over their implications and a discontent with the traditional labour–leisure choice model.

It has long been acknowledged that one reason why GDP per capita is a poor measure of economic well-being is that it does not recognize that leisure time has any value at all. If, as in the comparison of the United States with Germany or France, greater per capita GDP is obtained primarily from greater average working time, a comparison of economic well-being should measure both the cost of forgone individual leisure and the cost of the externality on the marginal utility of each individual's leisure as the decrease in the leisure time of everyone else impedes the feasibility of leisure time matches.

When (by increasing the availability of potential leisure matches) the choice of more leisure time by some individuals has a positive externality for other persons, there can be multiple equilibria in labour supply, in which the 'high work' equilibrium has unambiguously lower total utility. Societies which are better able to coordinate the level and timing of paid working hours may be better off in aggregate, because they enable their citizens to enjoy more satisfying social lives. To be specific, the leisure externality hypothesis suggests that Americans may work more hours than Europeans partly because they are more likely to have less satisfying social lives – because other Americans are also working more hours – and that they are worse off as a result.

Moreover, if authors such as Putnam (1993, 2000) and the OECD (2001) are correct in stressing the dependence of social capital on

associational life and the importance of social capital for social and economic development, the costs of a high-work/low-social life equilibrium may be substantial, in terms of market income as well as utility. Knack and Keefer (1997) are representative of an empirical literature which argues that localities with an active civic society and associational life (and more generally a dense network of social ties among individuals, and a high level of trust) have higher growth rates of GDP per capita. This relationship has been argued to be due to a number of possible influences: for example lower transactions costs in capital, labour and product markets, more effective governance, lower costs of crime, labour conflict and political uncertainty, better health outcomes, and so on (see Osberg 2003b). Whatever the channel of influence, it suggests that, although working longer hours may accelerate growth in GDP per capita in the short run, both income and social life may suffer in the longer run. There may be some wisdom in the old saying that: 'All work and no play makes Jack a dull boy.'

See Also

- ▶ External Economies
- ▶ Labour Supply
- ▶ Social Capital
- ▶ Time Use

Bibliography

Alesina, A., Glaeser, E. and Sacerdote, B. 2005. *Work and leisure in the U.S. and Europe: Why so different?* Working paper no. 11278. Cambridge, MA: NBER.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Bell, L., and R. Freeman. 2001. The incentive for working hard: Explaining hours worked differences in the US and Germany. *Labour Economics* 8: 181–202.

Corneo, G. 2005. Work and television. *European Journal of Political Economy* 21: 99–113.

Di Tella, R., R. MacCulloch, and A. Oswald. 2003. The macroeconomics of happiness. *The Review of Economics and Statistics* 85: 809–827.

Dickinson, D. 1999. An experimental examination of labor supply and work intensities. *Journal of Labor Economics* 17: 638–670.

Frey, B., and A. Stutzer. 2002. *Happiness and economics: How the economy and institutions affect well-being*. Princeton: Princeton University Press.

Hamermesh, D. 1998. When we work. *American Economic Review* 88: 321–325.

Hamermesh, D. 2002. Timing, togetherness and time windfalls. *Journal of Population Economics* 15: 601–623.

Jenkins, S., and L. Osberg. 2005. Nobody to play with? The implications of leisure co-ordination. In *The economics of time use*, ed. D. Hamermesh and G. Pfann. Amsterdam: Elsevier.

Juster, T., and F. Stafford, eds. 1985. *Time, goods, and well-being*. Ann Arbor: Institute for Social Research, University of Michigan.

Knack, S., and P. Keefer. 1997. Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics* 112: 1251–1288.

Kuhn, T. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Marshall, A. 1920. *Principles of economics*. 8th ed, 1961. London: Macmillan.

OECD (Organization for Economic Co-operation and Development). 2001. *The well-being of nations: The role of human and social capital*. Paris: OECD.

Osberg, L. 2003a. Understanding growth and inequality trends: The role of labour supply in the U.S.A. and Germany. *Canadian Public Policy* 29 (Supplement 1): S163–S183.

Osberg, L., ed. 2003b. *The economic implications of social cohesion*. Toronto: University of Toronto Press.

Putnam, R. 1993. *Making democracy work: Civic traditions in modern Italy*. Princeton: Princeton University Press.

Putnam, R. 2000. *Bowling alone: The collapse and revival of the American community*. New York: Simon and Schuster.

Winston, G. 1987. Leisure. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. Basingstoke: Palgrave.

Leisure Class

F. Stanković

Abstract

The idea of a leisure class was popularized by Thorstein Veblen, whose *Theory of the Leisure Class* (1899) developed the social categories of pecuniary competition, conspicuous leisure and conspicuous consumption. Bukharin's

Economic Theory of the Leisure Class (1919) argued that marginal utility theory was the theoretical expression of the class of rentiers who had been eliminated from the process of production and were interested only in disposing of their incomes. In *The Age of Uncertainty* (1977) J.K. Galbraith argued for the continuing relevance of Veblen's analysis. Modern sociologists, however, show little interest in the idea of a leisure class.

Keywords

Absentee ownership; Bukharin, N.; Class; Conspicuous consumption; Conspicuous leisure; Galbraith, J. K.; Labour; Labour theory of value; Leisure; Leisure class; Marginal utility theory; Marxism; Private ownership; Rentiers; Social status; Socialism; Subjective value theory; Veblen, T.

JEL Classifications

D6

This term became popular after Thorstein Veblen's book, *The Theory of the Leisure Class* (1899). In that book the author gives a historical and socio-economic explanation of the development of that wealthy class in the society of his time whose main characteristic was leisure. By 'leisure' Veblen means the non-productive spending of time which originates from a sense of the worthlessness of productive work and from the need to show pecuniary ability to afford a life of idleness. The basic social categories of Veblen's theory of the leisure class are pecuniary competition, conspicuous leisure and conspicuous consumption.

The leisure class is an old institution. The emergence of a leisure class coincides, according to Veblen, with the beginning of ownership. These two institutions (leisure class and ownership) are different aspects of the same general facts of social structure. The conditions for the appearance of the leisure class as a permanent form are: (1) the community has to be of a predatory type; and (2) the means of life must be affordable in a relatively easy way, so that part of the population can be liberated from routine work. It is necessary to make a proper distinction between the leisure and the labouring class. For the lower classes, as Veblen explains, since labour is their accepted and only mode of life, an emulative pride in a reputation for efficiency in their work becomes the only emulation that is open to them. For the 'superior pecuniary class' the most imperative secondary demand for emulation is the abstention from productive work. It is not sufficient to possess wealth or power. It becomes important to show that you have no need to do productive work. From the days of the Greek philosophers to the present a life of leisure is, as Veblen says, in a great part of secondary and derivative value (1936, p. 231).

Leisure does not usually bring about a material product; the result takes the form of non-material goods. Good examples of occupations that members of the leisure class choose to pursue are the knowledge of dead languages and the occult sciences, of correct spelling, of the various forms of music and other household arts, fashionable dress, furniture, games, sports, dogs and racehorses. Elegant speech shows the level of a speaker's emancipation from productive work. The leisure class is not interested in technological innovations and it is an obstruction to social and economic progress.

Veblen's critique of 'absentee ownership' is the next step in his analysis of modern civilization. The members of the leisure class do not want to have any connection with a production process and they leave the managing and guiding of this process to so-called 'captains of industry'. The last group is for Veblen the only positive power interested in technological development.

Veblen's critique of ownership and his opinion that the modern type of ownership is not compatible with industrial efficiency has been and still is unacceptable to many of his fellow economists. Like Marx, Max Weber and Karl Polanyi, Veblen had demonstrated the importance of studying primitive economies for general economic history

and the interaction between economics and the society in general.

Veblen's leisure class seems very much like Marx's ruling capitalist class. Both of them see the bitter struggle between capital and labour. But they differ very much according to the methods of solving this contradiction. Marx's solution is, as is well known, a socialist revolution. In *The Theory of the Leisure Class* Veblen was not explicit about this issue. He was not sure about the end of this struggle, but he was very positive about the existence of the struggle. This goes very well with his Darwinian philosophy. A kind of trade unionism could be closer to Darwinism and is more acceptable for Veblen than neo-Hegelianism or Marxism.

Although his theory is based on Darwinian philosophy and although he argues for evolutionary socialism, in some of his later writings he was closer to the attitude that certain radical social movements could be the solution for breaking with old social institutions. In his search for emancipation of man, Veblen, together with other great humanists, had to become a socialist. His analysis of private ownership and the leisure class necessarily pushed him into this direction.

Nikolai Bukharin's *Economic Theory of the Leisure Class* (1919) is quite different from Veblen's book. This is not an economic account of the conditions which give rise to the existence of a leisure class in the manner of Veblen. The leisure class is not the central category in this work, as the title may suggest. Bukharin mentions the leisure class in the context of his critique from a Marxist point of view of the theory of marginal utility, especially that of the Austrian school. He gives no reference to Veblen. Bourgeois political economy, according to Bukharin, seeks to justify the capitalist system and therefore it loses its scientific role, contrary to the Marxist theory which claims its general validity precisely for the reason that it is the theoretical expression of the most advanced class – the working class.

In the critique of marginal utility theory Bukharin points out that this theory is the theoretical expression of the class of rentiers who have been eliminated from the process of production and are interested in disposing of their income from holdings of securities and bonds only. This marginal utility theory is, according to Bukharin, unhistorical: it starts with consumption not the production process. His critique of the logic and the method of subjective value theory is settled in direct confrontation with the labour theory of value.

J.K. Galbraith refers to Veblen's work, suggesting that the concepts 'conspicuous leisure' and 'conspicuous consumption' are still of significance (Galbraith 1977). In the United States, as Galbraith explains, class as described by Veblen still exists: the members of the leisure class are still buying their social status.

It might have been expected that contemporary sociologists would have been more concerned with the concept of the leisure class. But this is not the case. They are very much occupied with leisure itself: it is common for sociologists to define leisure as the portion of time which remains when time for work and the basic requirements for existence have been satisfied, and this issue is related to the problem of how to spend leisure time. But this is a different problem and does not have a substantial relation to the problem of the leisure class.

See Also

- ▶ Bukharin, Nikolai Ivanovitch (1888–1938)
- ▶ Conspicuous Consumption
- ▶ Veblen, Thorstein Bunde (1857–1929)

Bibliography

Bukharin, N. 1919. *The economic theory of the leisure class*. New York: Monthly Review Press. 1972. (Originally published in Russian and first published in English in 1927).

Galbraith, J.K. 1977. *The age of uncertainty*. Boston: Houghton Mifflin.

Veblen, T. 1899. *The theory of the leisure class*. London: George Allen & Unwin.

Veblen, T. 1936. What Veblen taught. In *Selected writings of Thorstein Veblen*, ed. W.C. Mitchell. New York: Viking Press.

Lenin, Vladimir Ilyich [Ulyanov] (1870–1924)

Meghnad Desai

Keywords

Capitalism; Falling rate of profit; Hobson, J. A.; Imperialism; Lenin. V. I.; Marx's analysis of capitalist production; New Economic Policy (USSR); Planning; Socialism; State capitalism

JEL Classifications

B31

Vladimir Ilyich Ulyanov, who wrote and gained fame under the pseudonym Lenin, was born in April 1870, the. second son of a Russian provincial official in Simbirsk (now Ulyanovsk). After the arrest and execution of his elder brother Alexander, in 1887 for alleged terrorist activity, Lenin became increasingly active in political study groups at Kazan, Samara and St Petersburg. He came to identify himself with the Marxist rather than the populist (Narodniki) stream in these study groups. He played an active part in the early theoretical debates between these two streams on the future course of Russia's economic and political development. At the time of the founding of the Russian Social Democratic Labour Party (RSDLP) in 1898, he was already known as its best young theorist. A split in the RSDLP took place in 1902 and Lenin became identified as the leader of the majority (Bolshevik) faction. He spent much of the early years of the 20th century in exile in London, Paris and Zurich. He returned to Russia in April 1917 after the February Revolution had initiated the post-tsarist phase of Russian politics. Lenin, unlike his fellow party members, correctly foresaw the instability of the political situation in which an unelected liberal democratic cabinet uneasily shared power with the federation of popularly elected factory committees (Soviets). He launched the Bolsheviks on a strategy of revolutionary rejection of the government and a platform of peace in the World War at any price. His analysis proved correct when in November 1917 the Bolsheviks won a majority in the All Russian Congress of Soviets and took power. Lenin led the communist government from that day until illness forced his withdrawal from active politics in March 1923. He died in January 1924.

Lenin's economic writing is extensive, comprising books, pamphlets, newspaper articles and occasional speeches (see Desai (1986), for a full bibliography). His contributions can be placed under three headings: analysis of Russia's capitalist development in the period 1880–1900; the analysis of the developments in world capitalism in the period 1900–1916, where his concept of imperialism as a form of monopoly capitalism was an innovation; and lastly as a Marxist policy maker during the period 1917–23.

The Development of Capitalism in Russia

Lenin's book of this title published in 1898 is a substantial piece of work which traces the growth of commercial relations and specialization in agriculture leading to an erosion of the traditional communal forms. On the industrial side, Russia's late arrival entailed an active role for the tsarist state in fostering industrialization and an influx of foreign capital to finance the development. This meant that Russia, although a newly industrializing country in the 1890s, had a larger proportion of its industrial labour force in large factories than older industrialized countries like Britain. Lenin saw these as predictable consequences of rapid capitalist growth which made any going back to pre-capitalist communal forms of village organization impossible. The growth of large factories also meant concentration of workers in a few places, facilitating their combination in trade union activities. These economic circumstances – the growth of commercial relations in the countryside and of concentrations of the urban proletariat – dictated for Lenin the political strategy of a socialist party which hoped to

win power by mass organization. Lenin's theory of the development of the democratic political movement follows the economic stages quite closely. In this sense he can be said to have developed an economic framework for a Marxist political theory. *The Development of Capitalism in Russia* is even to this day the only comprehensive economic history of a country from a Marxist perspective.

Imperialism, the Highest Stage of Capitalism

In 1916, Lenin wrote his well-known economic pamphlet of this title. The background was provided by the First World War, which had broken out two years previously with enthusiastic participation by the working people of various combatant nations and the connivance of the socialist parties. The 'betrayal' by the workers and their political leaders was one factor in Lenin's urge to explain these events. The second urge was perhaps provided by a desire to integrate the facts of a war into a Marxist theory of the long-run development and eventual breakdown of capitalism.

Marx had predicted a tendency for the rate of profit to fall as capitalist development proceeded. Among the forces which may counteract this tendency was an increasing concentration in industry and the emergence of larger industrial units. In 1907, Hilferding in his *Finance Capital* had provided a theory and empirical evidence for the increasing integration of bank finance and industrial capital. The formation of trusts and cartels was helped by banks willing to finance mergers and controlling and interlocking equity holdings. Marxist economists saw the 20th century as entering a monopoly phase of capitalism in contrast to the competitive phase that Marx had written about.

Lenin's achievement is to add to the Marx–Hilferding account an international economic and political element. One part of his theory came from Hobson's *Imperialism.* As an underconsumptionist, Hobson linked the fight over African and Asian territory in the last decades of the 19th century among European nations to the search for outlets for surplus which could not be sold at home. Hobson took the view that this imperial search was irrational. Lenin, as a Marxist, saw the irrationality as a systematic functional element in a world of monopoly capital economies each of which was trying to stave off the falling rate of profit by exporting. The battle for markets could not however take place in a politically neutral context as envisaged by competitive economic theory. Large cartels and monopolies gave a few leading bankers and industrialists influence with the political governments of their country. The battle for markets thus became a struggle between developed capitalist nations for territory. It was the struggle for territory as a surrogate for markets which led to military confrontation between the major industrial nations and hence war. War was not however predicted to be a satisfactory solution to the problem of markets or of profitability. It was likely in Lenin's view to be the harbinger of proletarian uprising against the system in these countries which would end it.

Thus Lenin blends international political developments into a Marxian theory of capitalist development. Imperialism in Lenin's definition is the entire set of unequal economic relations between capitalist countries – between rival mature capitalist countries fighting for markets as well as between mature countries and developing economies which become their markets. Formal political control by one nation over another is not a necessary element in Lenin's view of imperialism. Although immensely influential in the interwar years due to Comintern orthodoxy, this theory has come under some attack recently (Warren 1980). It lacks a coherent analytical theory of how monopoly capital differs from competitive capitalism and its empirical predictions proved only temporarily true when a series of political uprisings took place in Europe after the First World War. These uprisings did not mature into a full-scale collapse of capitalism, which continues many decades after Lenin foresaw its highest phase as having been achieved.

Socialist Economic Policy

As the first Marxist to lead a government, Lenin had to formulate practical economic policy. Given the notorious lack of discussion of socialist economic policy in Marx's writings, Lenin had to improvise. Two notions stand out as his distinctive contribution to this area. First, in his description of the post-revolutionary Russia as a transitional state from capitalism to socialism. During this transition, state capitalism was seen by Lenin as an advance upon private capitalism in as much as the political state was not a capitalist one but a workers' state. Lenin used the wartime German economic organization as the ideal of a fully integrated single economic unit which a planned socialist economy could beneficially emulate. Second, in the return to normality after the Civil War – in his pamphlet 'The Tax in Kind' – Lenin sketched a theory for the role of trade in reviving economic activity. The key was to move from a forced requisition of food surpluses to a policy of tax in kind and encouraging exchange. A revival of agriculture was required for an industrial revival but the terms of trade between the two sectors was a crucial policy variable in this respect. Trade is seen as an antidote to economic bureaucracy in this pamphlet. It was this pamphlet that inaugurated the New Economic Policy which could be said to have lasted from 1921 to 1929.

Selected Works

1898. The development of capitalism in Russia. *In Collected works*, Vol. 3.
1916. Imperialism; the highest stage of capitalism. *In Collected works*, Vol. 22.
1921. The tax in kind. *In Collected works*, Vol. 32.
1960–70. *Collected works,* 45 vols. Trans. of the 4th enlarged Russian edn. Moscow: Progress Publishers.

Bibliography

Desai, M. 1986. *Lenin as an economist*. London: Lawrence &Wishart.
Warren, B. 1980. *Imperialism: Pioneer of capitalism*. London: New Left Books.

Bibliographic Addendum

Studies of aspects of Lenin's life and thought continue to be produced because of his importance in world history. Within this massive literature, valuable studies of his ideas include N. Harding, *Lenin's political thought,* 2 vols, New York: St. Martins Press, 1977 and 1981; and N. Harding, *Leninism.* Durham: Duke University Press, 1996. An extensive general biography is R. Service, R. *Lenin: A political life,* 3 vols, Bloomington: Indiana University Press, 1985, 1991, and 1995. This massive study is synthesized and updated in R. Service, *Lenin*, Cambridge, MA: Harvard University Press, 2000.

Leontief Paradox

Edward E. Leamer

Abstract

Using 1947 US input–output tables and data on exports and imports, Leontief (1953) found, to the surprise of the profession, that the capital per worker of US exports was less than the capital per worker of US import substitutes. The response to this empirical 'paradox' was the formulation of theory that might explain why a capital abundant country had labour-intensive exports. These were the first (confused) steps in an ongoing process of making the theory and data conform sufficiently to enable us comfortably to claim to understand the basis for international trade.

Keywords

Factor content of trade; Factor mobility; Heckscher–Ohlin–Samuelson trade model; Leontief paradox

JEL Classifications

F1

The Heckscher–Ohlin–Samuelson (HOS) model of international trade with two factors of production and two commodities implies that a country will export the commodity that is produced

intensively with the relatively abundant factor. Leontief (1953) discovered, to the surprise of the profession, that 1947 US exports were more labour-intensive than US imports in the sense that the capital per man required to produce a $1 million of exports was less than the capital per man required to produce a $1 million in import substitutes. This seemed to conflict sharply with the presupposition that the USA was abundant in capital compared with labour. Leontief's finding was so startling that it has been called a 'paradox', even though the result amounted to at most a single contradiction of the theory and even though no alternative model could be said to conform better with the facts.

Leontief's finding preceded and apparently stimulated a search of great breadth and intensity for a new theory of trade that could account for his result. It is in fact difficult to find another empirical result that has had as great an impact on the intellectual development of the discipline. Among the explanations of the finding are: (*a*) high productivity of US workers; (*b*) capital-biased consumption; (*c*) factor-intensity reversals; (*d*) tariffs; (*e*) abundance of natural resources; (*f*) abundance of human capital; (*g*) technological differences. These developments are surveyed in Chacholiades (1978, pp. 298–306).

It is surprising in retrospect that no one thought to examine the theoretical foundation for Leontief's inference that the factor content of US trade revealed the United States to be scarce in capital compared with labour, though a clear theory of the factor content of trade was not laid out until Vanek (1968). Vanek's model of the factor content of trade was first used in an overlooked article by Williams (1970) to criticize Leontief's inference. The very simple theoretical foundation for the Leontief calculation was clearly laid out in Leamer (1980), which shows that Leontief's data in fact reveal the United States to have been abundant in capital compared with labour.

Theoretical relationships that can serve as a foundation for studying the relative factor abundance revealed by international trade are the Heckscher–Ohlin–Vanek equations. These equations are derived from the simple identity that net exports of the services of a factor f are the difference between home supply and home demand: $T_f = X_f - M_f = S_f - D_f$, where T_f is the amount of factor *f* embodied in net exports, X_f is the amount of factor f required to produce the exported commodities, M_f is the amount required to produce the imported commodities, S_f is the domestic supply and D_f is the domestic demand. This identity is given empirical content by assuming identical homothetic tastes which implies that domestic demand for factor *f* is proportional to world supply, $D_f = sW_f$, where W_f is the world supply and *s* is the country's share of world consumption. With the use of this assumption, the net export equation can be written as

$$T_f/S_f = 1 - s\left(W_f/S_f\right).$$

In words, net exports as a share of domestic supply is positively related to factor abundance defined as the share of the world's total supply S_f/W_f. Accordingly, the relative scarcity of the factors is revealed by the ordering of the net export ratios T_f/S_f. Leamer (1980) shows that, although the net export of both capital and labour services were positive in 1947, the share of domestic supply of capital that was exported exceeded the share of labour exported, and consequently the United States was revealed by trade to be relatively abundant in capital compared with labour. In addition, Leamer (1980) shows that Leontief's finding that the exports were more capital intensive than imports is compatible with either ordering of factor abundance.

This fully resolves the apparent paradoxical ordering of capital and labour abundance, but a new problem arises. Brecher and Choudhri (1982) note that, if net exports are positive, the overall consumption share *s* must be less than the abundance ratio S_f/W. If trade is balanced, the consumption share is the ratio of home to world GNP, $s = \text{GNP}/\text{GNP}_w$. The inequality $S_f/W_f > s = GNP/GNP_w$ can be rewritten as $GNP_w/W_f > GNP/S_f$. Thus the United States is revealed by its positive net exports of labour services embodied in commodities to have had a per-capita GNP that is less than the rest of the world. Even after adjusting for the trade surplus, this is impossible to square with the facts. Another way of

expressing this new paradox is that the positive export of labour services reveals that labour is abundant compared with other resources on the average since the consumption share s is an average of the abundance ratios.

It is ironic that this is one of the few empirical findings that can be said to have had a decided impact on the course of the profession and at the same time is based on a simple conceptual misunderstanding. The error that is implicit in Leontief's paradox is the use of an intuitive but false theorem which states that the ordering of capital per man in exports compared with imports reveals the relative abundance of capital and labour. This is true for the simple two-good model, but it is not the case for a multi-commodity reality. There is a lesson to be learned from this experience.

Empirical work requires a fully articulated theoretical foundation. Intuition alone is not enough.

Although the precise form that Leontief's calculations took is inappropriate, the calculation of flows of factor services embodied in trade remains an interesting activity since these flows can be used to form a proper test of the Heckscer–Ohlin–Samuelson theorem and since the net effect of trade on the demand for factors of production can be an important input into trade policy that is intended to affect the distribution of income.

As it turns out, measurements of 1967 factor contents of trade reported in Bowen, Leamer and Sveikauskas (1987) rather badly violate the HOS model, thus reinvigorating the message of the Leontief paradox: there is something wrong with this model. One thing that is wrong is emphasized by Trefler's (1995) title: 'The Case of the Missing Trade'. Given the world's apparent unequal geographic distribution of capital, labour and land, the HOS model suggests that there should be much more trade than actually occurs. Trefler's solution to this puzzle is to allow in the model both home bias in consumption and also international productivity differences (for example, the United States is not so labour-scarce when allowance is made for the intensity of work). Also, Conway (2002) finds problems with the measurement of factor scarcity and calls for the model to include factor-specific differences in domestic factor mobility. It seems likely that we have not seen the end of the search for a model that most fully explains the nature of international trade.

See Also

- ▶ Heckscher–Ohlin Trade Theory
- ▶ Input–Output Analysis

Bibliography

Bowen, H., E. Leamer, and L. Sveikauskas. 1987. Multicountry, multifactor tests of the factor abundance theory. *American Economic Review* 77: 791–809.

Brecher, R., and E. Choudhri. 1982. The Leontief paradox, continued. *Journal of Political Economy* 90: 820–823.

Chacholiades, M. 1978. *International trade theory and policy*. New York: McGraw-Hill.

Conway, P. 2002. The case of the missing trade and other mysteries: Comment. *American Economic Review* 92: 394–404.

Leamer, E. 1980. The Leontief paradox, reconsidered. *Journal of Political Economy* 88: 495–503.

Leontief, W. 1953. Domestic production and foreign trade: The American capital position re-examined. *Proceedings of the American Philosophical Society* 97: 332–349. Reprinted in *Readings in international trade*, ed. H. Johnson and R. Caves. Homewood: R.D. Irwin. 1968.

Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85: 1029–1046.

Vanek, J. 1968. The factor proportions theory: The N-factor case. *Kyklos* 21: 749–756.

Williams, J. 1970. The resource content in international trade. *Canadian Journal of Economics* 3: 111–122.

Leontief, Wassily (1906–1999)

R. Dorfman

Keywords

Cobweb cycles; Concentration; Foreign aid; Implicit theorizing; Index numbers; Input–output analysis; Leontief paradox; Leontief, W.

JEL Classifications
B31

Wassily Leontief was born on 5 August 1906 in St Petersburg, the only child of an academic family. He studied first at the University of Leningrad, earning the degree of 'Learned Economist' in 1925, and then at the University of Berlin (Ph. D., 1928). While working on his doctorate, he was appointed a research economist at the University of Kiel, where he remained for about three years, with a year out to serve as adviser to the Chinese Ministry of Railways in Nanking.

In 1931 he went to the United States to join the staff of the National Bureau of Economic Research, but after only a few months accepted an appointment at Harvard University, where he remained for the following 44 years. During those years he attained worldwide eminence, particularly for the invention and application of input–output analysis. Prominent among the honours he received during those years were election as President of the American Economic Association in 1970 and the Nobel Memorial Prize in Economics in 1973. In 1975 he accepted a chair at New York University, where he spent the remainder of his career.

Leontief had an exceptionally strong training in mathematics and a marked flair for mathematical and geometric reasoning. These qualities were displayed in his earliest papers, in the late 1920s and early 1930s, in which he applied his technical talents to a variety of topics including the estimation of elasticities of supply and demand, the measurement of industrial concentration, the use of indifference maps at a time when they were still novelties to explain patterns of international trade in a two-commodity, two-country model, analysis of the conditions under which cobweb cycles would converge or would expand explosively, and several others. These papers established his reputation as an economic theorist of first rank.

During this same period, he struck a theme that he was to emphasize repeatedly throughout his career: the thesis that economic concepts were meaningless and misleading unless they could be observed and measured. Thus, in 1936 he studied the significance of index numbers that purported to measure composite concepts such as the aggregate output of an economy or the general price level, and the following year published his famous diatribe against 'implicit theorizing', that is, explaining phenomena by introducing ill-defined concepts (the economist's version of Molière's doctor who attributed the effect of sleeping potions to their dormative propensities). Eleven years later, he returned to the measurement of aggregates much more profoundly and fruitfully in his 'Introduction to a Theory of the Internal Structure of Functional Relationships', which developed the mathematical conditions in which a single aggregate or index could replace a mass of detailed data without loss of information. And much later he devoted his presidential address to the American Economic Association to decrying 'Theoretical Assumptions and Nonobserved Facts' (1971).

These two characteristics – adroitness at mathematical expression and analysis and insistence that theoretical concepts be implementable – congealed in Leontief's major achievement, the invention, development, and application of input–output analysis. As a purely theoretical construct, input-output analysis had a long genealogy before Leontief began his work on it, around 1933. In the 18th century, François Quesnay used his *Tableau économique* to illustrate the relationships between agriculture and other sectors of the economy. A hundred years later, Marx demonstrated the relationships between the capital-goods and consumers' goods departments of an economy by a very similar two-sector table. The most important predecessor, however, was Walras's formulation of the general equilibrium of an economy, which employed a concept that is very similar to Leontief's input–output coefficients. In addition, as Leontief discovered after input-output analysis was well known, H.E. Bray had published essentially the same equations in 1922, and R. Remak had discovered them again in 1929.

The algebraic theory of input–output analysis had been explored by a number of late 19th-century algebraists, particularly by E. Frobenius

and O. Perron, for whom the basic theorems have come to be named. All of these preceding theories expressed fundamental, abstract theoretical concepts; none could be used to specify the relationships among the sectors of an actual economy.

But throughout his career, Leontief has insisted that the task of a theorist only begins with the proposal of a well-formulated theory; the central task is to show that the theory can be applied to real economies, that it leads to interesting predictions about the behaviour of those economies, and that those predictions can be checked and found to be reasonably accurate. This radically operational point of view led Leontief to his critical contribution: the perception that the coefficients that express the relationships among the sectors of an economy can be estimated statistically, and that they are sufficiently stable so that they can be used in comparative static analyses to give quantitative estimates of the effects of different economic policies, taking into account their reverberations throughout the economy along with their effects on the industries affected in the first instance.

It is almost impossible now to appreciate the task of confirming these conjectures in the early 1930s. Input–output computations depend on inverting large matrices; the most powerful computing machines in existence then were punch-card machines that could multiply, after a fashion, but could not divide. Solving a half-dozen simultaneous linear equations was a formidable calculation; Leontief envisaged systems that numbered in the hundreds.

Input–output analysis also required data of an unfamiliar type – coefficients specifying the amounts of various raw materials and intermediate goods required per unit of product in each sector. The US Census of Manufactures included many of these coefficients, but by no means all. The remainder had to be compiled laboriously from trade journals and scattered sources.

Furthermore, the underlying assumption of the method, that the input–output coefficients remained essentially constant for substantial periods, was hard to reconcile with one of the main tenets of the theory of production – that factors of production were substituted for one another quite sensitively in response to price changes.

Beginning around 1933–4, Leontief concentrated on overcoming these difficulties by compiling coefficients for a 44-sector input–output table – about 2000 coefficients – and making plans for their analysis. Since the solution of 44 simultaneous equations was far beyond the realm of the possible, the 44 sectors were consolidated into a scant ten for computational purposes. To check on the stability of the coefficients, tables were to be compiled for 1919 and 1929.

The first result of this study, 'Quantitative Input and Output Relations in the Economic System of the United States', appeared in 1936. Its centrepiece was a 41-sector input–output table for the United States in 1919, presenting the intersectoral flow coefficients along with sources and methods of estimation. The next year, Leontief published 'Interrelation of Prices, Output, Savings and Investment'. In the interim, he had made contact with Professor John B. Wilbur of the Massachusetts Institute of Technology, who had just invented an analog computer that could solve systems of up to nine linear equations. Accordingly, Leontief aggregated his 41- sector table into ten sectors and used Wilbur's computer to calculate the inverse. This was the first Leontief-inverse ever computed, and probably the first use of a large computer in economics or other social science.

By 1941, a parallel 41-sector table had been compiled for 1929 and the inverse of a ten-sector aggregation of it had been computed. The two tables were presented and compared in Leontief's first monograph, *The Structure of American Economy, 1919–1929.* The comparisons were intended to test whether the input coefficients were stable enough to yield useful empirical predictions. The comparisons were indecisive, in part for lack of a clear standard for judging the stability of the estimated coefficients.

The monograph did establish, however, that it was feasible to compile the raw data needed for an input–output table and to compute coefficients and an inverse table that appeared to make good economic sense. The importance of such tables for economic planning was recognized almost

immediately. Within a few years, the US Bureau of Labour Statistics, with Leontief as a consultant, constructed a 400-sector table for projecting post-war employment by major industries, and the method was being applied all over the world for constructing economic development plans.

Leontief remained in the forefront of these developments. By 1944 he had calculated a table of input coefficients for 1939, comparable with the earlier two tables, and found a satisfactory degree of stability for most of the coefficients extending over two decades. Using this up-to-date table, he published a sequence of three important papers in the *Quarterly Journal of Economics* for 1944 and 1946 exemplifying the use of input–output analysis for estimating the effects of exogenous disturbances on output, employment, wages, and prices in individual sectors.

In 1948, Leontief established the Harvard Economic Research Project as a centre for applying and extending input–output analysis. He became director of the Project, and headed it for the next 25 years. He was particularly active in developing interregional input–output analysis and in introducing capital–coefficient matrices to derive the investment implications of changes in final demand and, thereby, to use input-output analysis to generate growth paths as well as static equilibriums of economic systems.

This work led to two books, *The Structure of American Economy 1919–1939*, in 1951, and *Studies in the Structure of the American Economy,* in 1953, as well as several international conferences and a score of papers and articles. Probably the most striking discovery of this period of work has come to be called 'the Leontief paradox', the finding that, when indirect as well as direct input requirements are taken into account, American exports are more labour-intensive and less capital-intensive than American imports, although the United States is exceptionally well endowed with capital and has exceptionally high real wages.

Leontief and the staff of the Harvard Economic Research Project devised and implemented numerous other applications of input–output analysis. They included estimates of the inflationary impact of wage settlements, calculations of the direct and indirect effects of armament expenditures on the individual sectors of the economy, and methods for projecting the growth-paths of the sectors in a developing economy and for estimating capital requirements for economic development.

In the middle 1970s, Leontief became persuaded that, while competitive markets might guide an economy to a socially efficient equilibrium if given sufficient time, the process would be likely to be very protracted and unduly wasteful of mistakenly invested resources. Economic growth and efficient adjustments would be promoted better by establishing an economic planning board that would work out a number of detailed growth possibilities based on input–output analyses. The ultimate choice among these possibilities would be made by a political process. He advocated this type of indicative planning in a number of articles in *The New York Review of Books*, the *New York Times* 'op-ed page', and other general interest periodicals.

Leontief subsequently turned to the problems of worldwide economic growth, its environmental impact, its demands on the world's base of natural resources, and particularly on its implications for relations between the economies of the so-called First and Third Worlds. Under the sponsorship of the United Nations, he directed a study of the evolution of the world economy until the year 2000, based on a multiregional input-output model consisting of 15 regions, each comprised of 45 sectors, and linked by balanced trading relationships. This is, perhaps, the most ambitious input–output study yet undertaken. The results were published as *The Future of the World Economy* (1977). It found that, under a wide range of plausible assumptions, little progress would be made in closing the gap between the industrial and the developing regions unless current policies concerning international trade and finance were changed drastically in the directions of increased multinational aid and an increased flow of imports from the Third World to the First.

Leontief was a leader in improving the computational methods of economics, beginning with his use of Wilbur's analog equation solver in 1936. Subsequently he inverted input–output matrices on Howard Aiken's early Mark I and

Mark II computers, the immediate predecessors of the electronic computer. In the 1980s the very large matrices required by his world economic models led him to be the first economist to use the so-called supercomputers and to apply parallel-processors and other highly efficient methods of computation.

Throughout his career, Leontief took an active interest in the education of the next generation of scholars. While at Harvard he served for 11 years as chairman of the Society of Fellows, the foundation that provides three-year, duty-free fellowships to promising young scholars, to enable them to reside at Harvard and pursue whatever interests they choose. He delighted in presiding over the weekly dinner meetings of the Society and leading conversations that range over all the fields of interest represented at the table.

See Also

► Input–Output Analysis

Selected Works

Leontief's principal scientific contributions can be found in four volumes:

1941. *The structure of American economy, 1919–1929.* Cambridge, MA: Harvard University Press and later editions.

1953. (With others). *Studies in the structure of the American economy.* New York: Oxford University Press.

1966. *Essays in economics: Theories and theorizing.* New York: Oxford University Press, and later editions.

1977. *Essays in economics,* Vol. 2. White Plains: M.E. Sharpe.

Articles exemplifying Leontief's wide range of interests and activities can be found in:

Bulletin of the American Mathematical Society, April 1947.

New York Review of Books, 10 October 1968, 21 August 1969, 4 June 1970, 7 January 1971, 20 July 1972, 4 December 1980, 12 August 1982.

New York Times, op-ed page, 14 March 1974, 24 March 1977, 6 March 1979, 5 April 1981, 19 September 1983.

Scientific American, April 1961, September 1963, April 1965, September 1980.

Bibliography

A range of essays evaluating and extending Leontief's work may be found in.

Dietzenbacher, E., and M. Lahr. 2004. *Wassily Leontief and input–output economics*. Cambridge: Cambridge University Press.

Wood, J. 2001. *Wassily Leontief: Critical assessments*. London/New York: Routledge.

Lerner, Abba Ptachya (1905–1982)

T. Scitovsky

Keywords

Cost-push inflation; Demand-pull inflation; Distributional optimality; Excess-claims inflation; Excess-demand inflation; Expectational inflation; Expected and unexpected inflation; Factor-price equalization; Full employment; Functional finance; Incomes policy; Inflation; International trade theory; Keynes, J. M.; Lerner, A. P.; Low full employment; Marginal cost pricing; Market Anti-Inflation Plan (MAP); Market pricing; Market socialism; Monopoly; Natural rate of unemployment; Optimality; Optimum currency areas; Organization of Petroleum Exporting Countries (OPEC); Pareto efficiency; Perfect competition; Socialist free enterprise; Stagflation; Unemployment; Welfare economics

JEL Classifications

B31

Lerner was one of the last of the great non-mathematical economists and certainly one of

the most original, versatile and prolific members of the profession. Born in Rumania, raised from early childhood in the Jewish immigrant quarter of London's East End, he went to rabbinical school, started work at 16, working as tailor, capmaker, Hebrew School teacher, typesetter, and then founded his own printing shop. When that went bankrupt at the onset of the Great Depression, he enrolled as an evening student at the London School of Economics to find out the reason for his shop's failure. There, his outstanding logical faculties soon became evident and won him all the available prizes and fellowships, one of which took him to Cambridge to study with Keynes. He published many major articles already as an undergraduate, was appointed temporary assistant lecturer at the London School of Economics in 1935, assistant lecturer in 1936, and in 1937 a Rockefeller fellowship took him to the United States, where he remained, although his restlessness kept him from settling at any one university for more than a few years.

Lerner was a lifelong socialist, advocate of market pricing for its allocative efficiency, and believer in private enterprise, whose offer of private employment he considered an essential safeguard of individual freedom. That unusual combination of principles accounts for Lerner's loneliness and political isolation. In his economics, however, he knew how to reconcile those principles. His reconciliation of the first two made him into one of the founders (along with Oskar Lange) of the theory of market pricing in the decentralized socialist economy, and he sought to reconcile the first and third principles by advocating what he called socialist free enterprise: 'the freedom of both public and private enterprise to enter any industry on fair terms which, in each particular case, permit that form to prevail which serves the public best.'

Although Lerner's ambition was to improve the economy, not economics, he made many, often fundamental contributions to economic theory, mainly in the fields of welfare economics, international trade and macroeconomics but also in the theories of production, capital, monopoly, duopoly, spatial competition and index numbers. Furthermore, and hardly less important, he made generous use of his geometrical skill and genius for exposition in tidying up and clarifying other people's ideas. As a result, a number of important economic theorems and ideas, though first stated by others, became the profession's common property in Lerner's simpler and clearer formulations. An important example of that is the well-known rule that marginal cost pricing is a condition of welfare optimality. Another example is his definitive proof (Lerner 1936a) that in the two-country, two-commodity model, export and import duties have identical consequences if their proceeds are spent in the same way.

In welfare economics, one of his first articles (Lerner 1934a) not only introduced the notion that monopoly is a matter of degree, whose extent is best measured by the excess of price over marginal cost, but in the process also provided the first complete, comprehensive and clear statement and discussion of the nature and limitations of Pareto optimality, and of the equality between price and marginal cost and between price and marginal value product as necessary conditions of optimality. All that, along with Lerner's many papers on market pricing under socialism, was restated, elaborated and extended in his 1944 *The Economics of Control: Principles of Welfare Economics.*

That work, Lerner's best book, became and remains the most comprehensive non-mathematical text on welfare economics. Although written in the style of a handbook, with its propositions presented as rules for the planners and plant managers of a decentralized socialist economy to follow, the book is better described by the second than by the first half of its title. For most of those rules are nothing but the first-order conditions of optimality, presented with great care, clarity and completeness but without a hint at the practical obstacles in the way of putting them into actual practice. As a text on welfare economics, however, it is exceptionally meticulous and complete, it extends the scope of the welfare principle from resource allocation narrowly defined to taxation, macroeconomics and international trade and finance, and it contains the first logically based analysis of distributional optimality. Moreover, since a socialist economy, for Lerner, meant the use of private enterprise in some sectors, state-

owned plants in others, depending on which was the more efficient in each, his guidebook for socialist planners also discusses why and when perfect competition leads to optimality and why and when real-life competition falls short of being perfect.

In the field of international trade theory, Lerner derived Samuelson's celebrated factor-price equalization theorem 15 years before Samuelson in a 1933 unpublished seminar paper printed only 19 years later (Lerner 1952a). His elegant and ingenious resolution of a 19th-century controversy over the identity of import and export duties has already been mentioned; he devised (Lerner 1932, 1934b) the standard geometry of the two-country, two-commodity model, which is well known from a whole generation of textbooks; and he was the first to raise and deal with the question of 'optimum currency areas' in his 1944 *Economics of Control.*

Most of Lerner's innovations in microeconomics and international trade theory were so basic and so useful that they promptly became integral parts of every economist's standard equipment. That is why it is hard to appreciate, at this late stage, the striking originality and elegant simplicity of his logic. One gets a glimpse of that by looking at his almost unknown proposal of how to counter OPEC's raising of the price of oil (Lerner 1980a). He proposed the imposition of a variable import duty on oil (which he called extortion tax), whose level would always match the producer's profit margin, thereby rising and falling with the oil price and being higher on imports from high-priced and lower on those from low-priced producers. Since such a tariff would make consumers face much larger price changes than those decided upon by OPEC and much greater price differentials than those set by the different oil exporters, it would also make consumers' responses to those price changes and differentials correspondingly greater, thereby raising the price elasticity of demand for oil as it appears to producers. That would lower OPEC's monopoly power and so its profit maximizing monopoly price, and it would increase the rewards and the temptation for OPEC members to break up the coalition by defecting from it.

In macroeconomics, Lerner did as much as anyone to clarify, extend and popularize Keynes's *General Theory*; he was the first to recognize the inflationary implications of employment policies, the first to analyse in depth and in detail the causes and nature of inflation, and to propose a remedy for stagflation.

Lerner wrote the first article (1936b) to make Keynes's employment theory simple and generally intelligible, and in two short papers clarified Keynes's 'user cost' and 'marginal efficiency of capital' concepts (1943b, 1953). He wrote an interesting book (1951) to summarize and significantly extend Keynes's employment theory; he published an enlightening paper to explain the *General Theory*'s obscure Chapter 17 (1952b), thereby clearing up the complex role wage rigidity plays in rendering underemployment equilibrium possible; and he was the person best to elucidate the relation between macroeconomics and microeconomics by representing them as the two limiting cases of a more general type of economic analysis (1962).

Next to his work on welfare economics and international trade theory, Lerner's best known and most shockingly new contribution was his introduction of the idea of 'functional finance' (1943a; also restated in 1951, and in his 1944 *Economics of Control*), whose advocacy of Keynesian employment policies exposed the latter's logical implications and revolutionary nature. To careless readers, it also seemed like a wildly inflationary doctrine, although Lerner's concern over inflation and over the inflation effects of employment policies antedate everybody else's by many years.

Lerner's extensive work on inflation began with his distinguishing between low and high full employment (Lerner 1951). High full employment is that beyond which further demand expansion presses against supply limitations and creates overspending (demand-pull) inflation; low full employment is the employment level below which the price level is stable. Levels of employment between the low and high full-employment levels create administered (cost-push) inflation, owing to labour's excessive bargaining strength. His 'low full employment' therefore is a

forerunner (by 17 years) of Friedman's 'natural rate of unemployment'.

Lerner's theoretical papers on inflation contain many pioneering insights. One is his sharp analytic distinction between overspending or excess-demand inflation and administered or excess-claims inflation (1958, 1972), of which the former does, but the latter (according to him) does *not*, call for fiscal and/or monetary restraint. He later added a third category, expectational inflation (1972), which he also called defensive inflation to differentiate it from the aggressive nature of excess-claims inflation – arguing that incomes policy is effective against the former but ineffective against the latter. Another and well-known distinction which Lerner was the first to draw was that between expected and unexpected inflation (1949).

Since Lerner's heart was in reform, not in analytic niceties, his many discussions of inflation were just a preamble for working out a plan to control the main economic problem of his time, stagflation, that is, the combination of unemployment and inflation, which he considered characteristic of administered or excess-claims inflation. Restrictive policies were to him an inadmissible cure for that type of inflation, because he considered the creation of unemployment a prohibitive cost. Incomes policies he judged ineffective against all but expectational inflation, and he was too ardent a believer in the pricing mechanism to argue for wage and price controls. He wanted to stabilize the general price level without impeding the free movement of individual prices and wages. To accomplish that, he devised and, with David Colander's help, worked out in detail a scheme, called Market Anti-Inflation Plan, better known as MAP (1980b), for rationing the right of firms to raise the 'effective price' of their output, that is, the sum of profits and wages entering the price of their products (value added). The scheme would give every firm the right to increase its value added in the proportion of the estimated rise in the economy's overall productivity, but it would also allow them to sell their unused rights or the unused portion of their rights (in a market created for the purpose) to those other firms that want to increase their wages and/or profits (value added) in greater proportion.

Lerner developed his Market Anti-Inflation Plan gradually and published it at several stages and in several versions before it reached its final form in 1980. It was his last major contribution to economics and a fitting end to his career, because it well illustrates both the strengths and the weaknesses of his extraordinarily fertile and original mind. It is bold, elegant, ingenious and impeccably logical, with meticulous attention to every conceivable detail and exception, but combines those qualities with a slightly utopian flavour, all of which have characterized just about all of Lerner's many proposals for reform.

For the sheer novelty and stark logic of Lerner's arguments and policy proposals usually took people aback, but he was utterly unwilling and perhaps also unable to soften their impact in the interests of their easier acceptability. He was well aware of the reasons for the hostile reception of virtually all his recommendations but believed, with some justification, that, as time wore off their shocking novelty, they would become more acceptable and politically feasible. Lerner's MAP could well be the best remedy for stagflation but many less good remedies will first have to be tried and prove ineffective in order to render MAP politically acceptable.

Selected Works

1932. The diagrammatical representation of cost conditions in international trade. *Economica* 12: 346–356.

1934a. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1(June): 157–175.

1934b. The diagrammatical representation of demand conditions in international trade. *Economica,* NS 1: 319–334.

1936a. The symmetry between import and export taxes. *Economica,* NS 3: 306–313.

1936b. Mr. Keynes' 'General theory of employment, interest and money'. *International Labour Review* 34: 435–454.

1943a. Functional finance and the federal debt. *Social Research* 10(February): 38–51.
1943b. User cost and prime user cost. *American Economic Review* 33(March): 131–132.
1944. *The economics of control: Principles of welfare economics.* New York: Macmillan.
1949. The inflationary process: Some theoretical aspects. *Review of Economics and Statistics* 31(August): 193–200.
1951. *Economics of employment.* New York: McGraw-Hill.
1952a. Factor prices and international trade. *Economica,* NS 19: 1–15.
1952b. The essential properties of interest and money. *Quarterly Journal of Economics* 66(May): 172–193.
1953. On the marginal product of capital and the marginal efficiency of investment. *Journal of Political Economy* 61(February): 1–14.
1958. Inflationary depression and the regulation of administered prices. *Joint Economic Committee Print*, Conference on Economic Stability and Growth, March.
1962. Macro-economics and micro-economics. In *Logic, methodology and philosophy of science: Proceedings of the 1960 international congress,* ed. E. Nagel, P. Suppes and A. Tarski. Stanford: Stanford University Press.
1972. *Flation: Not inflation of prices, not deflation of jobs.* New York: Quadrangle Books.
1980a. OPEC – A plan – If you can't beat them, join them. *Atlantic Economic Journal* 8(3): 1–3.
1980b (with D.C. Colander.) *MAP, A market anti-inflation plan.* New York: Harcourt, Brace and Jovanovich.

Bibliography

For a representative collection of Lerner's best writings, see D.C. Colander, ed., *Selected economic writings of Abba P. Lerner,* New York: New York University Press, 1983. That volume contains most of the articles cited here and also has Lerner's complete bibliography.
For other, more detailed appraisals of Lerner's contribution to economics, see:
Samuelson, P.A. 1964. A.P. Lerner at sixty. *Review of Economic Studies* 31 (June): 169–178.
Scitovsky, T. 1984. Lerner's contribution to economics. *Journal of Economic Literature* 22: 1547–1571.
Sobel, I. 1979. Abba Lerner on employment and inflation: A post-Keynesian perspective. In *Essays in post-Keynesian inflation*, ed. J.H. Gapinski and C.E. Rockwood. Cambridge, MA: Ballinger.

Leroy-Beaulieu, Pierre-Paul (1843–1916)

R. F. Hébert

Keywords
Iron law of wages; Leroy-Beaulieu, P.-P.; Mathematical method; Public finance; Ricardian theory of rent

JEL Classifications
B31

French economist and journalist, Leroy-Beaulieu was born at Paris in 1843; he died there in 1916. His father was a Prefect and a Deputy under Louis-Philippe, his older brother a famous historian and a director of the Ecole des Sciences Politiques. His son Pierre, with whom he is sometimes confused, was also an economist. Initially trained in law, Paul Leroy-Beaulieu turned to economics in his early twenties, launching this new career with a prize-winning essay in 1867 on the effects of the moral and intellectual conditions of the working class on the rate of wages. Soon thereafter, he began collaborating on the *Revue des deux mondes,* and in 1871 he became editor of the *Journal des débats.* Two years later he founded the *Economiste française,* for which, as editor, he wrote weekly articles, missing only once in 43 years.

When Emile Boutmy established the Ecole Libre des Sciences Politiques in 1872, Leroy-Beaulieu accepted the chair of public finance. He later succeeded his father-in-law, Michel Chevalier, in the chair of political economy at the Collège de France. His ideas found wide exposure in countless journal articles and over a dozen

books. A member of the French Institute and of the American Philosophical Society, he also received honorary degrees from the universities of Cambridge, Edinburgh, Dublin and Bologna.

Leroy-Beaulieu belonged to the French Liberal School of individualism and free trade. His major work, the *Traité théorique et pratique d'économie politique* (1896) is largely an exposition of classical theory. However, he rejected the pessimistic conclusions of Ricardo and Malthus, having argued in his *Essai sur la répartition des richesses* (1881) that there was no factual basis to either the Ricardian theory of rent or the 'iron law of wages'. Moreover, he sought to defuse the population bomb by arguing that the progress of civilization must always bring a declining birth rate because the altered demands and increased expenditures that accompany it are incompatible with the duties and responsibilities of parentage. In value theory, he followed the marginal analysis of the Austrians. Even as Walras was proselytizing on its behalf, however, Leroy-Beaulieu reviled the mathematical method as 'pure delusion and a hollow mockery . . . [without] scientific foundation and . . . practical use'. Showing equally poor judgement, he rejected the demand curve on frivolous grounds.

Leroy-Beaulieu's most enduring work was his treatise on public finance (1877), an effort that examines both public revenues and public credit. The second volume of this work rose somewhat above the first, remaining authoritative well into the 20th century.

Selected Works

1877. *Traité de la science des finances*, 2 vols. Paris: Guillaumin.
1881. *Essai sur la répartition des richesses et sur la tendance à une moindre inégalité des conditions*. Paris: Guillaumin.
1890. *L'état moderne et ses functions*. Paris: Guillaumin.
1896. *Traité théorique et pratique d'économie politique*, 4 vols. Paris: Guillaumin.
1913. *La question de la population*. Paris: F. Alcan.

Bibliography

Pirou, G. 1925. *Les doctrines économiques en France depuis 1870*. Paris: A. Colin.
Stourm, R. 1917. Paul Leroy-Beaulieu. *Revue des deux mondes,* Series VI 38: 532–553.

Level Accounting

Francesco Caselli

Abstract

Level accounting (more recently known as development accounting) consists of a set of calculations whose purpose is to find the relative contributions of differences in inputs and differences in the efficiency with which inputs are used to cross-country differences in GDP. It is therefore the cross-country analogue of growth accounting.

Keywords

Development accounting; Growth accounting; Level accounting; Technical change; Total factor productivity

JEL Classifications

D4; D10

Suppose that country A is observed to produce more output than country B: is this because it employs a larger amount of labour, a larger amount of capital or a larger amount of some other input? Or because it somehow succeeds (or endeavours) to make more effective use of given inputs? Level accounting refers to a particular approach to attacking these questions. In this approach, one computes indices of the quantities of each input participating in production in different countries, as well as the shares of each input in total income. The contribution of inputs (or of a subset of the inputs) to differences in output is then given by a geometric average of the inputs,

with the shares acting as weights. The difference between the cross-country difference in output and the cross-country differences in inputs, a residual, is interpreted as a cross-country difference in the efficiency with which the inputs are employed, or in total factor productivity (TFP). Level accounting is therefore the cross-country analogue of growth accounting.

The earliest level-accounting exercises are a five-country study by Denison (1967) and a two-country comparison by Walters (1968). In the late 1970s Jorgenson and Nishimizu (1978) and Christensen et al. (1981) adapted the growth-accounting framework of Jorgenson's work with Griliches and Christensen to level comparisons between the United States and eight other advanced economies. They found substantial TFP differences.

More recently, level accounting has been a popular technique in addressing the sources of the enormous differences in income observed between the richest and poorest economies of the world (King and Levine 1994; Klenow and Rodriguez-Clare 1997; Hall and Jones 1999). This trend has caused several authors to begin referring to it as 'development accounting'. While details vary, a consensus emerging from the development-accounting literature is that observed inputs of labour and capital account for at best 50% of the observed variation in aggregate value added across a large sample (numbering about 100) of developed and developing countries. It is often argued that this evidence points to the need for developing countries to underemphasize saving and investment, and emphasize technical change and technology adoption.

Unfortunately, residual variation in development accounting poses at least as many problems of interpretation as residual variation in growth accounting. The problems are compounded by the appalling coarseness of the data. Instead of accounting for compositional differences amongst a large number of education, gender, race, and age categories, as mandated by the Jorgensonian framework, development accountants to date have mostly had to limit themselves to a rough correction for average years of schooling. Perhaps more importantly, instead of allowing for imperfect substitutability among different types of capital, again as prescribed by best accounting practice, measures of the capital stock are based on linear aggregation. Caselli and Wilson (2004) show that this could be a fatal flaw. Finally, most development-accounting exercises assume constant capital (and hence labour) shares across countries.

Creative improvements in the measurement of labour quality have recently been proposed by Weil (2007) and Jones and Schneider (2007). Weil proposes a way to account for differences in the productive capacity of the labour force caused by differences in health, while Jones and Schneider bring to bear cross-country differences in IQ. Both succeed in reducing residual variation considerably. These appear to be two (rare) instances where level accounting has introduced innovations that could potentially also be usefully incorporated into growth accounting, instead of the other way around.

Another recent extension of the development-accounting framework is due to Caselli and Coleman (2006), who show how to decompose the cross-country residual into differences in the efficiency with which different inputs are used. Caselli (2005) uses this technique to show that most differences in efficiency are differences in the efficiency with which labour is used. Caselli and Coleman (2006) further trace these differences to differences in the efficiency of skilled labour.

Cross-country level accounting can also be performed at the industry level, and indeed this seems a necessary step towards shedding light on the sources of large residual variation at the aggregate level. Conrad and Jorgenson (1985), and Jorgenson et al. (1987) presented industry-level productivity comparisons for the United States, Japan, and Germany. Despite the richness of their data they found surprisingly large TFP differences. The more recent development- accounting literature has only attempted an agriculture-nonagriculture decomposition. The most convincing effort to date is possibly due to Vollrath (2006), who appears to be able to eliminate a significant amount of residual variation in aggregate GDP by accounting for the allocation of factors across these two sectors.

See Also

► Growth Accounting

Bibliography

Caselli, F. 2005. Accounting for cross-country income differences. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Caselli, F., and J. Coleman. 2006. The world technology frontier. *American Economic Review* 96: 499–522.

Caselli, F., and D. Wilson. 2004. Importing technology. *Journal of Monetary Economics* 51: 1–32.

Christensen, L., D. Cummings, and D. Jorgenson. 1981. Relative productivity levels, 1947–1973: An international comparison. In *New developments in productivity measurement and analysis, studies in income and wealth*, vol. 41, ed. J. Kendrick and B. Vaccara. Chicago: University of Chicago Press.

Conrad, K., and D. Jorgenson. 1985. Sectoral productivity gaps between the United States, Japan, and Germany, 1960–1979. In *Probleme und Perspektiven der Wirtschaftlichen Entwicklung*, ed. H. Giersch. Berlin: Duncker and Humblot.

Denison, E. 1967. *Why growth rates differ?* Washington, DC: Brookings Institution.

Hall, R.E., and C.I. Jones. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114: 83–116.

Jones, G., and Schneider, J. 2007. IQ and productivity. Working paper, George Mason University.

Jorgenson, D., and M. Nishimizu. 1978. U.S. and Japanese economic growth, 1952–1974. *Economic Journal* 88: 707–26.

Jorgenson, D., M. Kuroda, and M. Nishimizu. 1987. Japan–U.S. industry-level productivity comparisons, 1960–1979. *Journal of the Japanese and International Economies* 1: 1–30.

King, R.G., and R. Levine. 1994. Capital fundamentalism, economic development, and economic growth. *Carnegie-Rochester Conference Series on Public Policy* 40: 259–92.

Klenow, P.J., and A. Rodriguez-Clare. 1997. The neoclassical revival in growth economics: Has it gone too far? In *NBER macroeconomics annual 1997*, ed. B.-S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Vollrath, D. 2006. *How important are dual economy effects for aggregate productivity?* Houston: University of Houston Press.

Walters, D. 1968. *Canadian income levels and growth: An international perspective*. Ottawa: Economic Council of Canada.

Weil, D. 2007. Accounting for the effect of health on economic growth. *Quarterly Journal of Economics* 122: 1265–1306.

Lewis, W. Arthur (1915–1991)

Ronald Findlay

Abstract

This article provides an outline of the career and contributions of W. Arthur Lewis, who was awarded the Nobel Prize for Economics in 1982. Born in 1915 on the Caribbean island of St Lucia, Lewis began his career at the London School of Economics before moving to Manchester University and then to Princeton University. While *The Theory of Economic Growth* (1955) and *Growth and Fluctuations 1870–1913* (1978) have both been regarded as classics since they were first published, it is the 1954 article, 'Economic Development with Unlimited Supplies of Labour', that will probably remain as Lewis's most famous and influential single contribution.

Keywords

Development economics; Dual economies; Lewis, W. A.; Nurkse, R.; Periphery; Surplus labour; Terms of trade

JEL Classifications

B31

W. Arthur Lewis was born on the island of St Lucia in the British West Indies on 23 January 1915. His early education was at St Mary's College on the island, where he completed a rigorous high school curriculum by the age of 14. This school is remarkable for having been attended not only by Lewis but also, 15 years later, by St Lucia's other Nobel Laureate, the poet Derek Walcott. A scholarship took Lewis to the London School of Economics in 1933, where he obtained a BA in Commerce with first class honours in 1937 and then went on to do a Ph.D. under the supervision of Arnold Plant, who incidentally was also the supervisor of Ronald Coase. In 1938 he was appointed as a junior member of the

faculty, the first black man to receive such a position in the history of the institution. His very active teaching at the LSE on a very broad range of subjects undoubtedly prepared him well for his future work on economic development. He moved to Manchester University in 1947, where he held the Stanley Jevons Chair, previously occupied by J.R. Hicks, and where he was himself to be succeeded by Harry Johnson. It was here that he did some of his most seminal work on development economics, the *Manchester School* article on 'Economic Development with Unlimited Supplies of Labor' (1954) and the treatise on *The Theory of Economic Growth* (1955). In the 1950s he was a senior official in agencies of the United Nations, and was for a time Vice Chancellor of the University of the West Indies. He went to Princeton in 1963, where he remained until his retirement in 1983: just as at the LSE, he was the first person of African descent ever to be appointed to the faculty. He held many parttime advisory positions with international organizations and governments in developing countries, particularly in West Africa and the Caribbean. He was awarded the Nobel Prize for Economics in 1979, together with T.W. Schultz, for their contributions to economic development. He died at his summer home in Barbados on 15 June 1991.

His earliest original research, including his Ph.D. thesis, was on the application of price theory to problems of industrial organization and public utilities. A number of studies published during the 1940s, such as 'The Two Part Tariff' (1941), 'Competition in Retailing' (1945), 'Fixed Costs' (1946), and other related topics, were brought together in a volume entitled *Overhead Costs,* published in 1949. Two other books published in the same year, based on his LSE lectures, were *Economic Survey 1919–1939* and *Principles of Economic Planning.* The first of these was an examination of the troubled economic history of the world economy in the interwar period, notable in particular for the way in which he linked together the experiences of the 'core' industrial countries with those of the primary producing 'periphery' of the world economy. The pessimism about the possibility of international trade to serve as a sustained 'engine of growth' for the developing countries, that has marked his subsequent writings on development economics down to his Nobel Prize Lecture in 1980 (entitled the 'The Slowing Down of the Engine of Growth'), can perhaps be traced to his study of the inter-war period, an interesting parallel with the case of Ragnar Nurkse, who also came to the study of development problems after writing his *International Currency Experience* on the breakdown of the international monetary system in the 1930s. The book on planning, though written at an introductory level, was a penetrating early examination of the problems of coordinating government intervention and the market in a mixed economy.

Lewis's most famous and influential contribution to economics is undoubtedly the 1954 paper on development with 'unlimited supplies' of labour. He presents a stylized model in which the typical poor country is divided into a 'traditional' and a 'modern' sector. The former consists of peasant agriculture as well as selfemployment of various sorts in urban areas, where the primary objective of economic activity is to maintain consumption. The 'modern' sector comprises commercial farming, plantations and mines and manufacturing, in all of which there is hired labour and profit is the motive for production organized by a class of capitalists and entrepreneurs. Lewis adopts a strictly classical viewpoint on two crucial features of his model. First, the real wage of unskilled labour in the modern sector is exogenously given, with employment and profits then being determined by the demand for labour corresponding to the fixed stock of capital in the short run.

The second classical feature is that the accumulation of capital is governed by saving out of profits. The process of economic development is viewed as the expansion of the modern relative to the traditional sector until such time as the 'surplus labour' pool in the traditional sector is drained and an integrated labour market emerges with a neoclassically determined equilibrium real wage, rising steadily over time as growth proceeds. The model as a whole thus has two distinct phases, an initial 'classical' one with a fixed real wage, that is the main focus of the analysis, and a subsequent 'neoclassical' one with a rising real wage. The concept of a 'dual economy' in the first phase of the model has generated considerable controversy

and an extensive polemical literature, to which references can be found in Findlay (1980), together with an appraisal, extensions and critique of the model itself. The most sophisticated and thorough theoretical defence of the dual economy and the associated notion of 'surplus labour' remains that provided by Sen (1966). The *Manchester School* in 2004 appropriately marked the 50th anniversary of the most celebrated article it ever published by a special issue, which contains a valuable survey of subsequent developments by Kirkpatrick and Barrientos (2004).

Another notable, but much less well-known, contribution of this seminal (1954) paper, in a neglected section on the open economy, is a model of the terms of trade between manufactures and primary products that is developed further, with empirical applications, in his (1969) Wicksell Lectures. The key idea is that the world price of manufactures, relative to the prices of tropical products such as coffee, tea, sugar, rubber and jute, is determined by the relative opportunity costs of labour in food production. Thus the Pittsburgh steel worker's wage is governed by the Kansas farmer's productivity, while the Brazilian coffee plantation wage is determined by the much lower productivity of peasant subsistence agriculture, which explains why a unit of steel in the world market commands so many more units of coffee. Since the transformation curves between steel and food and coffee and food are assumed to be linear, demand only determines quantities produced, consumed and traded, not relative prices, exactly as in the approach of the classical economists. Lewis applied this model in a very imaginative way to illuminate several key aspects of the history of the world economy in his last major work, *Growth and Fluctuations 1870–1913,* published in (1978). This volume extended his examination of the world economy in the inter-war period in *Economic Survey 1919–1939* back to the 'golden age' of globalization from 1870 to 1913, and is a deeply original piece of theoretical, statistical and historical research in the manner of Schumpeter and Kuznets. Both volumes are still essential reading for any serious student of the evolution of the world economy.

The reader can find an extensive collection of Lewis's articles and shorter monographs in the volume edited by Mark Gersovitz (1983). A measure of his influence on the field of development economics can be gathered from the volume of essays in his honour edited by Gersovitz and others (1982). Robert L. Tignor (2006) is a very valuable account of the life and inspiring achievements of this great pioneer of development economics, rightly drawing attention to the stoic courage and steely resolution with which he confronted and overcame the racial prejudice that was so virulent even in Western academic circles during his early career. The effect of these experiences may have made him appear to many as reserved, aloof and 'prickly' but to all who knew him well he was always kind, courteous and considerate, with a puckish sense of humour. The writer Pico Iyer (1997) described Derek Walcott, the other Nobel Laureate of St Lucia, as a 'Tropical Classical' because of the deep influence of Homer and other classical authors on his poetry. The designation fits Arthur Lewis admirably as well, and not only because of the influence of Ricardo and other classical authors on his economics.

See Also

► Dual Economies

Selected Works

1941. The two part tariff. *Economica* NS8: 240–70.
1945. Competition in retail trade. *Economica* NS12: 202–34.
1946. Fixed costs. *Economica* NS13: 231–58.
1949. *Economic survey 1919–1939.* London: Allen & Unwin.
1949. *Overhead costs.* London: Allen & Unwin.
1949. *The principles of economic planning.* London: Allen & Unwin.
1954. Economic development with unlimited supplies of labor. *Manchester School* 22: 139–91.

1955. *The theory of economic growth*. London: Allen & Unwin.
1969. *Aspects of tropical trade 1883–1965*. Wicksell Lectures, Stockholm: Almqvist & Wiksell.
1978. *Growth and fluctuations 1870–1913*. London: Allen & Unwin.
1980. The slowing down of the engine of growth (Nobel Lecture). *American Economic Review* 70: 555–64.

Bibliography

Findlay, R. 1980. On W. Arthur Lewis' contributions to economics. *Scandinavian Journal of Economics* 82 (1): 62–76.
Gersovitz, M. 1983. *Selected economic writings of W. Arthur Lewis*. New York: New York University Press.
Gersovitz, M., C.F. Diaz-Alejandro, G. Ranis, and M.R. Rosenzweig. 1982. *The theory and experience of economic development: Essays in honour of W. Arthur Lewis*. London: George Allen & Unwin.
Iyer, P. 1997. *Tropical classical: Essays from several directions*. New York: Vintage.
Kirkpatrick, C., and A. Barrientos. 2004. The Lewis model after 50 years. *The Manchester School* 72: 679–690.
Nurkse, R. 1944. *International currency experience*. Geneva: League of Nations.
Sen, A.K. 1966. Peasants and dualism with or without surplus labor. *Journal of Political Economy* 74: 425–450.
Tignor, R.L. 2006. *W. Arthur Lewis and the birth of development economics*. Princeton: Princeton University Press.

Lexicographic Orderings

Charles Blackorby

Keywords
Interpersonal utility comparisons; Lexicographic orderings; Maximin; Social choice

JEL Classifications
O1

Lexicographic orderings are orderings in which certain elements of the space being ordered have been selected for special treatment. I begin with an example. Suppose an agent has an ordering over commodities *a* and b. Although he or she likes both *a* and b, any bundle which has more of *a* is *preferred* to any bundle which has less of *a*. Of course among bundles which have the same amount of *a*, bundles with more *b* are preferred to those with less. Thus, there are no trade-offs between *a* and *b* and each indifference set is a single point. The name 'lexicographic' comes from the way words are ordered in a dictionary, alphabetically by the first letter and then the second and so on.

Lexicographic orderings were known chiefly as simple examples of orderings which *could not* be represented by a continuous real-valued function; see Debreu (1954) for the first discussion of this issue in economics. It is, however, in social choice theory and welfare economics where these orderings have come to prominence. To demonstrate their role a lexicographic maximin rule (leximin) follows. Let $u = (u_1, \ldots, u_N)$ be an element of a Euclidean N-space where u_n is the utility of person n. In each possible state of the world, say $\bar{u}_1 = (\bar{u}_1, \ldots, \bar{u}_n)$, let $r(\bar{u})$ be the person who is the rth best off. For example, if $N = 3$ and $\bar{u} = (2, 7, 3)$ then $1(\bar{u}) = 2$ as person 2 has the highest utility, $2(\bar{u}) = 3$, and $3(\bar{u}) = 1$; ties are broken arbitrarily. An ordering R is a leximin rule if and only if for all $(u, \bar{u})$, $\bar{u}P\bar{u}$ if and only if there exists a k, $1 \leq k \leq N$, such that $\bar{u}_{k(u)} > \bar{u}_{k(u)}$ and for all $j > k$, $\bar{u}_{j(u)} = \bar{u}_{j(u)}$ where P is the strict preference relation, the asymmetric factor of R. That is, if the worst-off $N - k$ people have the same utility levels in $\bar{u}$ and $\bar{u}$ and the next worst-off person, k, is better off in $\bar{u}$ than in $\bar{u}$, then $\bar{u}$ is preferred to $\bar{u}$. Continuing the numerical example above let $\bar{u} = (2, 7, 2.5)$ so that $1(\bar{u}) = 2$, $2(\bar{u}) = 3$, and $3(\bar{u}) = 1$. Then $k = 2$, $\bar{u}_{2(u)} = 3 > \bar{u}_{2(u)}$ and $\bar{u}_{3(u)} = 2 = \bar{u}_{3(u)}$ hence $\bar{u}P\bar{u}$.

It is important to notice that if each person's utility function were subjected to the same increasing transformation the above ordering would not change. This is a case where utility is ordinally measurable but fully comparable as

levels of utility can be compared across individuals. That the leximin rule satisfies all of the original axioms of Arrow (1951, 1963) except for the comparability of levels of utility was first worked out by D'Aspremont and Gevers (1977).

Other types of lexicographic orderings appear frequently in social choice theory; see Sen (1986, section 6).

See Also

► Orderings

References

Arrow, K.J. 1951. *Social choice and individual values*. 2nd ed. New York: Wiley. 1963.

D'Aspremont, C., and L. Gevers. 1977. Equity and the informational basis of collective choice. *Review of Economic Studies* 44: 199–209.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision processes*, ed. R. Thrall, C. Coombs, and R. Davis. New York: Wiley.

Sen, A.K. 1986. Social choice theory. In *Handbook of mathematical economics*, ed. K. Arrow and M. Intriligator, vol. 3. Amsterdam: North-Holland.

Lexis, Wilhelm (1837–1914)

S. L. Zabell

German economist and statistician; born at Eschweiler, Germany, 17 July 1837; died at Gottingen, Germany, 24 August 1914.

Although Lexis's initial training was in mathematics and the natural sciences (he obtained a degree in mathematics, wrote a thesis on analytical mechanics, and for a brief period did research in Bunsen's laboratory), he almost immediately turned to the social sciences. After graduating from the University of Bonn in 1859, Lexis went to Paris in 1861 to study economics, and in 1870 published his first important work, a study of French export policy. Over the next decade and a half Lexis gained recognition, holding a succession of academic positions: professor of economics at Strasbourg (1872); professor of geography, ethnology, and statistics at Dorpat (1874); professor of economics at Freiburg (1876) and Breslau (1884); and professor of political science at Gottingen (1887), where he remained until his death 27 years later during the opening days of World War I.

Lexis's most important and lasting research contributions were to statistics and demography, but he wrote extensively (and primarily) about economics and actuarial science. He was a founding member in 1872 of the *Verein für Sozialpolitik* (part professional association, part pressure group, and part research organization), a coeditor of the *Handwörterbuch der Staatswissenschaften* (the leading German economic encyclopedia), and, after 1891, editor of the *Jahrbücher für Nationalökonomie und Statistik*. His seminars in insurance at Gottingen, taught from 1895 until his death, were the first of their kind in Germany.

Reflecting his scientific background and training, Lexis's work in economics displayed a concern for grounding economic theories in empirical quantitative reality. He was a critic of the Austrian and Lausanne schools, and in particular the work of Menger, Auspitz and Lieben. While agreeing with Gossen in some areas, Lexis differed with Gossen's law of equalization of marginal utilities, and was sceptical about the value of marginal utility theory.

Lexis's most lasting contribution to demography was the the simple but useful *Lexis diagram* (Lexis 1875, p. 302). This was a graphical representation of lifetable data which facilitated computation. In the Lexis diagram. the abscissa (t) represents time and the ordinate (x) represents age. The lifeline of an individual is a straight line at 45° to both axes, beginning at the point corresponding to date of birth and zero age, and terminating at the point corresponding to date of death and age at death. The diagram is still used today; see, e.g., Pressat (1972, *passim*).

The Lexis Ratio

Lexis's interest in statistics arose out of problems he had encountered in economics, demography, and sociology. Here too his role was that of sceptic. The work of Quetelet and his followers often made the implicit but unsupported assumption of temporal or spatial homogeneity in sampling from human populations. In order to test this assumption, Lexis proposed a test statistic Q, known today as the *Lexis ratio*. This statistic contrasts the observed variability in the data with the variability to be expected under the hypothesis of homogeneity; its use thus corresponds to the present-day variance test for the homogeneity of two or more binomial proportions $p_1, p_2, \ldots, p_n$. The Lexis ratio is note-worthy as one of the earliest instances of the use of the analysis of variance, although Lexis had been partially anticipated in this by Bienaymé, Campbell, and Dormoy.

If the populations being sampled are referred to as *strata*, and the individuals sampled within a population as *types*, then three cases may be distinguished, depending on whether the sampling probabilities are 1) constant across both strata and type (*Bernoulli sampling*); 2) constant within a stratum but vary from strata to strata (*Lexian sampling*); 3) constant for a given type across strata, but vary from type to type within a stratum (*Poisson sampling*); this terminology is due to Charlier. The Lexis ratio was scaled so that the *theoretical dispersion coefficient* $D = \sqrt{E(Q)}$ would have value $D = 1$ (*normal dispersion*) in the first case, value $D > 1$ (*supernormal dispersion*) in the second, and value $D < 1$ in the last (*subnormal dispersion*). In actual practice Q was replaced by an *empirical dispersion coefficient* Q^*.

The work of Lexis was later developed by his student, the Russian emigré Ladislaus von Bortkiewicz, but never received a fully satisfactory mathematical foundation until the work of the Russian mathematicians Chuprov and Markov from 1910 to 1920. Using a modified Lexis ratio $L = \{(mn-1)/n\,(m-1)\}Q^*$, where m denotes the number of strata, and n the number of types, Chuprov proved in 1916 that $E[L] = 1$, and soon after Markov was able to show that Var /Var $[L] \leq 2/(m-1)$ for $m \geq 5$. Markov's work culminated in 1920 with a completely rigorous proof, using his method of moments, that the asymptotic distribution of L is (up to a scaling factor) chi-squared on $m - 1$ degrees of freedom. This was an impressive technical achievement for its time and one which, unknown to either Karl Pearson or R.A. Fisher, had already decided their later, celebrated, degrees of freedom controversy.

Despite its importance, the statistical work of Lexis was largely ignored in England, save by Edgeworth, whose writings provided, as Keynes observed, 'for nearly forty years past, on this as on other matters where the realms of Statistics and Probability overlap, almost the only connecting link between English and continental thought' (Keynes 1921, pp. 394–5). A particularly spectacular illustration of this intellectual gulf was the Lexis ratio itself: although algebraically equivalent, up to multiplicative constant, to Pearson's chi-squared statistic, this equivalence went unnoticed until 1924, when it was noted by Fisher (1928, p. 807; [1925] 1970, p. 80). For further biographical information, a detailed bibliography, and a critical assessment of Lexis's contributions to economic theory, see Heiss (1978, pp. 507–12); see also L. von Bortkiewicz (1915); and Oldenburg (1933). For useful historical background, see Anthony Oberschall (1965).

For the Lexis diagram, see Wilhelm Lexis (1875, p. 302). Modern discussions of the diagram include Nathan Keyfitz (1968, pp. 9–11); and Roland Pressat (1972, pp. 15ff).

For discussion of Lexis's dispersion theory in the context of 19th century statistics, see Stephen M. Stigler. (1986, ch. 6) and Theodore M. Porter (1986, pp. 240–55). The subsequent history of the Lexis ratio is discussed by C.C. Heyde and E. Seneta (1977, pp. 49–58). For the work of Chuprov and Markov, see K.O. Ondar (1981). For an extensive account of the impact of the Lexis ratio on statistical theory, see Rainald K. Bauer (1955).

J.V. Uspensky (1937, pp. 212–30) gives a detailed mathematical treatment of the Lexis ratio, including many of the results of Chuprov and Markov. Other useful accounts include those of Arne Fisher (1915), chs. 10–12 (which contains many interesting empirical examples); Georg

Polya (1919); Julian Lowell Coolidge (1924, pp. 66–73); H.L. Rietz (1927, ch. 6); and F.N. David (1949). Of particular interest is Keynes's discussion of the Lexis ratio in his *Treatise on Probability* (1921, ch. 32).

For the connection between the Lexis ratio and Pearson's chi-squared statistic, see R.A. Fisher (1928) and R.A. Fisher ([1925], 1970).

Selected Works

1875. *Einleitung in die Theorie der Bevölkerungsstatistik*. Strasbourg: Trübner.

Bibliography

Bauer, R.K. 1955. Die Lexissche Dispersionstheorie in ihren Beziehungen zur modernen statistichen Methodenlehre, inbesondere zur Streuunganalyse. *Mitteilungsblatt für mathematische Statistik und ihre Anwendungsgebiete* 7: 25–45.

Bortkiewicz, L. Von. 1915. *Bulletin of the International Statistical Institute* 20: 328–332.

Coolidge, J.L. 1924. *An introduction to mathematical probability*. Oxford: Clarendon Press.

David, F.N. 1949. *Probability theory for statistical methods*. Cambridge: Cambridge University Press.

Fisher, A. 1915. *The mathematical theory of probabilities*. New York: Macmillan.

Fisher, R.A. 1925. *Statistical methods for research workers*. 14th ed, 1970. New York: Hafner Press.

Fisher, R.A. 1928. On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematicians, Toronto 1924*: 805–813.

Heiss, K.-P. 1978. Lexis, Wilhelm. In *International Encyclopedia of Statistics*, ed. W. Kruskal and J. Tanur. New York: Free Press.

Heyde, C.C., and E. Seneta. 1977. *I.J. Bienaymé: Statistical theory anticipated*. New York: Springer.

Keyfitz, N. 1968. *Introduction to the mathematics of population*. Reading: Addison-Wesley.

Keynes, J.M. 1921. *Treatise on probability*. London: Macmillan.

Oberschall, A. 1965. *Empirical social research in Germany, 1840–1914*. Paris: Mouton & Co..

Oldenburg, K. 1933. Lexis, Wilhelm. In *Encyclopedia of the social sciences*, Vol. 9. London: Macmillan.

Ondar, O., ed. 1981. *The correspondence between A.A. Markov and A.A. Chuprov on the theory of probability and mathematical statistics*. New York: Springer.

Polya, G. 1919. Anschauliche und elementare Darstellung der Lexisschen Dispersionstheorie. *Zeitschrift für schweizerische Statistik und Volkswirtschaft* 55: 121–140.

Porter, T.M. 1986. *The rise of statistical thinking 1820–1900*. Princeton: Princeton University Press.

Pressat, R. 1972. *Demographic analysis: Methods, results, applications*. Chicago: Aldine.

Rietz, H.L. 1927. *Mathematical statistics*. Washington, DC: Mathematical Association of America.

Stigler, S.M. 1986. *The history of statistics*. Cambridge, MA: Harvard University Press.

Uspensky, J.V. 1937. *Introduction to mathematical probability*. New York: McGraw-Hill.

Liability for Accidents

Steven Shavell

Abstract

Legal liability for accidents determines the circumstances under which injurers must compensate victims for harm. The effects of liability on incentives to reduce risk, on risk-bearing and insurance (both direct coverage for victims and liability coverage for injurers), and on administrative expenses are considered. Liability is also compared with other methods of controlling harmful activities, notably, with regulation and corrective taxation.

Keywords

Accident insurance; Contributory negligence; Corrective taxes; Damages; Due care; Judgment-proof problem; Liability for accidents; Liability insurance; Moral hazard; Negligence rule; Product liability; Risk aversion; Safety regulation; Strict liability

JEL Classifications

D6; D8; H8; K13; L5

Legal liability for accidents governs the circumstances under which parties who cause harm to others must compensate them. There are two basic rules of liability. Under *strict liability*, an injurer must always pay a victim for harm due to an accident that he causes. Under the *negligence*

rule, an injurer must pay for harm caused only when he is found negligent, that is, only when his level of care was less than a standard of care chosen by the courts, often referred to as *due care*. (There are various versions of these rules that depend on victims' care, as will be discussed.) In fact, the negligence rule is the dominant form of liability; strict liability is reserved mainly for certain especially dangerous activities (such as the use of explosives). The amount that a liable injurer must pay a victim is known as *damages*.

Our discussion of liability begins by examining how liability rules create incentives to reduce risk. The allocation of risk and insurance is next addressed, and, following that, the factor of administrative costs. Then a number of topics are reviewed. Comprehensive economic treatments of accident liability are presented in Landes and Posner (1987) and Shavell (1987); an early, insightful informal, economically oriented treatment of liability is presented in Calabresi (1970). Empirical literature is surveyed in Kessler and Rubinfeld (2007) and is not considered here.

Incentives

In order to focus on liability and incentives to reduce risk, we assume in this section that parties are risk neutral. Further, we suppose that there are two classes of parties – injurers and victims – who do not have a contractual relationship. For example, injurers might be drivers and victims pedestrians, or injurers might be polluting firms and victims affected residents.

Unilateral Accidents and the Level of Care

Here we suppose that injurers alone can reduce risk by choosing a level of *care*. Let x be expenditures on care (or the money value of effort) and $p(x)$ be the probability of an accident that causes harm h, where p is declining in x. Assume that the social objective is to minimize total expected costs, $x + p(x)h$, and let x^* denote the optimal x.

Under strict liability, injurers pay damages equal to h whenever an accident occurs, and they naturally bear the cost of care x. Thus, they minimize $x + p(x)h$; accordingly, they choose x^*.

Under the negligence rule, suppose that the due care level $\hat{x}$ is set equal to x^*, meaning that an injurer who causes harm will have to pay h if $x < x^*$ but will not have to pay anything if $x \geq x^*$. Then the injurer will choose x^*: he will not choose $x > x^*$, for that will cost him more and he escapes liability by choosing merely x^*; he will not choose $x < x^*$, for then he will be liable (in which case the analysis of strict liability shows that he would not choose $x < x^*$).

Thus, under both forms of liability, injurers are led to take optimal care, as first shown in Brown (1973). Note that under the negligence rule courts need to be able to calculate optimal care x^* and to observe actual care x, in addition to observing harm. Under strict liability courts need only to observe harm.

It should also be noticed that, under the negligence rule with due care $\hat{x}$ equal to x^*, negligence is never found, because injurers are induced to be non-negligent. Findings of negligence may occur, however, under a variety of modifications of our assumptions. Courts might make errors in observing injurers' care, so that an injurer whose true x is at least x^* might mistakenly be found negligent because his observed level of care is below x^*. Similarly, courts might err in calculating x^* and thus might set due care $\hat{x}$ above x^*. If so, an injurer who chooses x^* would be found negligent (even though care is accurately observed) because $\hat{x}$ exceeds x^*. As emphasized by Craswell and Calfee (1986), error in the negligence determination leads injurers to choose incorrect levels of care, and under some assumptions, to take excessive care in order to reduce the risk of being found negligent by mistake. Other explanations for findings of negligence are that individuals may not know x^* and thus take too little care, the judgment-proof problem (see below), which may lead individuals to choose to be negligent, and the inability of individuals to control their behaviour perfectly at every moment or of firms to control their employees.

Bilateral Accidents and Levels of Care

We now assume that victims also choose a level of care y, that the probability of an accident is $p(x,y)$ and is declining in both variables, that the social

goal is to minimize $x + y + p(x,y)h$, and that the optimal levels of care x^* and y^* are positive.

Under strict liability, injurers' incentives are optimal conditional on victims' level of care, but victims have no incentive to take care because they are fully compensated for their losses. However, the usual strict liability rule that applies in bilateral situations is strict liability with a defense of *contributory negligence,* meaning that an injurer is liable for harm only if the victim's level of care was not negligent, that is, his level of care was at least his due care level $\hat{y}$. If victims' due care level is y^*, then it is a unique equilibrium for both injurers and victims to act optimally: victims choose y^* in order to avoid having to bear their losses, and injurers choose x^* since they will be liable because victims are non-negligent.

Under the negligence rule, optimal behaviour is also the unique equilibrium. Injurers choose x^* to avoid being liable, and, since victims therefore bear their losses, they choose y^*. Two other variants of the negligence rule are negligence with the defence of contributory negligence (under which a negligent injurer is liable only if the victim is not negligent) and the comparative negligence rule (under which a negligent injurer is only partially liable if the victim is also negligent). These rules also induce optimal behaviour.

Thus, all of the negligence rules, and strict liability with the defence of contributory negligence, support optimal care, on the assumption due care levels are chosen optimally. Courts need to be able to calculate optimal care levels for at least one party under any of the rules, and in general this requires knowledge of the function $p(x, y)$. The main conclusions of this section were first proved by Brown (1973) (see also Diamond 1974, for closely related results).

Unilateral Accidents, Level of Care, and Level of Activity

Now let us reconsider unilateral accidents, allowing for injurers to choose their level of *activity z,* which is interpreted as the (continuously variable) number of times they engage in their activity (or, if injurers are firms, the scale of their output). Let $b(z)$ be the benefit from the activity, and assume the social object is to maximize $b(z) - z(x + p(x)h)$; here $x + p(x)h$ is assumed to be the cost of care and expected harm each time an injurer engages in his activity. Let x^* and z^* be optimal values. Note that, as before, x^* minimizes $x + p(x)h$, and that z^* satisfies $b'(z) = x^* + p(x^*)h$, the marginal benefit from the activity equals the marginal social cost, comprising the sum of the cost of optimal care and expected accident losses.

Under strict liability, injurers choose both the level of care and the level of activity optimally, as their objective is the social objective.

Under the negligence rule, injurers choose optimal care x^* as before, but their activity is socially excessive. Because an injurer escapes liability by taking care of x^*, he chooses z to maximize $b(z) - zx^*$, so that z satisfies $b'(z) = x^*$. The injurer's cost of raising his activity level is only his cost of care x^*, which is less than the social cost, as that also includes $p(x^*)h$. The excessive level of activity under the negligence rule is more important the larger is the expected harm $p(x^*)h$ from the activity.

The failure of the negligence rule to control the level of activity arises because negligence is defined here (and for the most part in reality) in terms of care alone. A justification for this assumption is that courts might face informational difficulties were they to include the activity level in the definition of negligence. The problem with the activity level under the negligence rule is applicable to any aspect of behaviour that would be difficult to incorporate into the negligence standard (including, for example, research and development activity). The distinction between levels of care and levels of activity was developed in Shavell (1980).

Bilateral Accidents, Levels of Care, and Levels of Activity

If we consider levels of care and of activity for both injurers and victims, then none of the liability rules that we have considered leads to full optimality (on the assumption that activity levels are unobservable). The reason that full optimality cannot be achieved is in essence that injurers must bear full accident losses to induce them to choose the right level of their activity, but this

means that victims will not choose the optimal level of their activity.

Risk-Bearing and Insurance

We next examine the implications of risk aversion and the role of insurance in the liability system (see Shavell 1982a). A number of general points may be made.

First, the socially optimal resolution of the accident problem now involves not only the reduction of losses from accidents but also the protection of risk-averse parties against risk. Risk bearing is relevant for two reasons: not only because potential victims may face the risk of accident losses, but also because potential injurers may face the risk of liability. The former risk can be mitigated through accident insurance, and the latter through liability insurance.

Second, the incentives associated with liability do not function in the direct way discussed in the previous section, but instead are mediated by the terms of insurance policies. To illustrate, consider strict liability in the unilateral accident model with care alone variable, and assume that insurance is sold at actuarially fair rates. If injurers are risk averse and liability insurers can observe their levels of care, injurers will purchase full liability insurance coverage and their premiums will depend on their level of care; their premiums will equal $p(x)h$. Thus, injurers will want to minimize their costs of care plus premiums, or $x + p(x)h$, so they will choose the optimal level of care x^*. In this instance, liability insurance eliminates risk for injurers, and the situation reduces to the previously analysed risk-neutral case.

If, however, liability insurers cannot observe levels of care, ownership of full coverage could create severe moral hazard, so would not be purchased. Instead, as is known from the theory of insurance, the typical amount of coverage purchased will be partial, for that leaves injurers with an incentive to reduce risk. In this case, therefore, the liability rule results in some direct incentive to take care because injurers are left bearing some risk after their purchase of liability insurance, but their level of care tends to be less than first best.

This last situation, in which liability insurance dilutes incentives, leads to a third point, concerning the question whether the sale of liability insurance is socially desirable. (We note that, because of fears about incentives, the sale of liability insurance was delayed for decades in many countries and that it was not allowed in the Soviet Union; further, in the United States liability insurance is sometimes forbidden against certain types of liability, such as against punitive damages.) The answer to the question is that, even though it may dilute incentives, sale of liability insurance is socially desirable, at least in basic models of accidents and some variations of them. In the case just considered, for example, injurers are made better off by the presence of liability insurance, as they choose to purchase it, and victims are indifferent to its purchase by injurers because victims are fully compensated for any harm suffered. This argument must be modified in other cases, such as when the damages injurers pay are less than harm because injurers are judgment-proof.

Fourth, consider how the comparison between strict liability and the negligence rule is affected by risk bearing. The immediate effect of strict liability is to shift the risk of loss from victims to injurers, whereas the immediate effect of the negligence rule is to leave the risk on victims (as injurers tend to act non-negligently). However, the presence of insurance means that victims and injurers can substantially shield themselves from risk, attenuating the relevance of risk bearing for the comparison of strict liability and negligence.

Finally, the presence of insurance implies that the liability system cannot be justified primarily as a means of compensating risk-averse victims against loss. Rather, the justification for the liability system must lie in significant part in the incentives that it creates to reduce risk. To amplify, although both the liability system and the insurance system can compensate victims, the liability system is much more expensive than the insurance system (see the next section). Accordingly, were

there no social need to create incentives to reduce risk, it would be best to dispense with the liability system and to rely on insurance to accomplish compensation.

Administrative Costs

The administrative costs of the liability system are the legal and other costs (notably the time of litigants) involved in bringing suit and resolving it through settlement or trial. These costs are substantial; a number of estimates suggest that, on average, administrative costs of a dollar or more are incurred for every dollar that a victim receives through the liability system (Shavell 2004, p. 281).

Strict Liability Versus Negligence

The factor of administrative costs affects the comparison of liability rules. On one hand, we would expect the volume of cases – and thus administrative costs – to be higher under strict liability than under the negligence rule. On the other hand, given that there is a case, we would anticipate administrative costs to be higher under the negligence rule because due care will be at issue. Hence, it is not clear which liability rule is administratively cheaper.

Social Desirability of the Liability System and Private Motives to Sue

The existence and the surprisingly high magnitude of administrative costs raise rather sharply the question whether the liability system is socially worthwhile. Moreover, the private motive to sue is not in alignment with the social reasons for using the liability system. First, the private benefit of suit is the amount of money that would be obtained from it, whereas the social benefit is the deterrence that would be created. Second, the private cost of suit is the victim's cost, whereas the social cost includes also the injurer's and the state's cost. These differences give rise to the possibility of socially excessive or socially insufficient suit. To illustrate the former, suppose that care has no effect on the accident probability, so that it is socially undesirable for suit to be brought. Yet under strict liability a victim will bring suit as long as his cost is less than the harm suffered, so the volume of litigation activity could be high. To illustrate the possibility of socially inadequate suit, suppose that an expenditure on care of only one hundredth of harm will eliminate the possibility of otherwise certain harm, and suppose also that the magnitude of harm is less than the cost of suit. Then no suit will be brought. However, it would be desirable for victims to have an incentive to bring suit, for that would induce care to be taken, and, since no harm would then occur, no suit would ever occur. The private versus the social motive to make use of the legal system was first developed in Shavell (1982b, 1997); see also Polinsky and Rubinfeld (1988).

Topics

Damages

Under strict liability, damages must equal harm h for incentives to be optimal. Under the negligence rule, however, damages higher than h also would induce injurers to take optimal care of x^*. Higher damages will increase the incentive to be non-negligent; they will not lead injurers to take excessive care because injurers can escape liability merely by taking care of x^*. But when there is uncertainty in the negligence determination, damages higher than h may lead to problems of excessive care.

Damages exceeding h are desirable if injurers sometimes escape liability, as when injurers may be hard to identify (the origin of pollution may be difficult to trace). If the probability of liability for harm is q, then, if damages are raised to $(1/q)\,h$, expected liability will be h. Thus, the more likely an injurer is to escape liability, the higher should be damages. On these points and others about punitive damages, see Cooter (1989) and Polinsky and Shavell (1998).

Causation

A fundamental principle of liability law is that a party cannot be held liable unless he was the cause

of losses. For example, if cancer occurs in an area where a firm has polluted, the firm will be liable only for the cancer that it caused, not for cancer due to other carcinogens. This principle is necessary to achieve social efficiency under strict liability, because otherwise incentives would be distorted. Socially desirable production might be rendered unprofitable if the firm were held responsible for all cases of cancer. Under the negligence rule, restricting liability to accidents caused by an actor may be less important than under strict liability: if negligent actors were held liable for harms they did not cause, they would only have greater reason to act non-negligently. On causation and incentives, see Calabresi (1975), Kahan (1989), and Shavell (1987).

Judgment-Proof Problem

The possibility that injurers may not be able to pay in full for the harm they cause is known as the judgment-proof problem and is of substantial importance, for individuals and firms often cause harms significantly exceeding their assets. When injurers are unable to pay fully for the harm they may cause, their incentives to reduce risk are inadequate, and their incentives to engage in risky activities excessive. Policy responses to the judgment-proof problem include vicarious liability (imposed on a party who has some control over the judgment-proof party), minimum asset requirements for participation in harmful activities, safety regulation, and criminal liability. On the judgment-proof problem and responses to it, see Kornhauser (1982), Pitchford (1995), Shavell (1986, 2005), and Sykes (1984).

Product Liability

When victims are customers of firms, the role of liability in providing incentives may be attenuated or even non-existent. If customers have perfect knowledge of product risks, then they will pay less for risky products, and incentives to reduce risk will be optimal without liability. If, however, customer knowledge of risk is imperfect, liability is potentially useful in reducing risk. In the latter case, a question of interest is whether court-determined liability or market-determined liability, namely, warranties, is likely to be better, on which see Priest (1981), Rubin (1993), and Spence (1977).

Liability Versus Other Means of Controlling Risk

Liability is only one method of controlling harm-causing behaviour; safety regulation and corrective taxes are among the alternatives. Liability harnesses the information that victims have about the occurrence of harm, and thus may be advantageous when victims, rather than the state, naturally observe how harm comes about; whereas when harm-causing behaviour and its occurrence requires state effort to be ascertained, regulation and taxation may be advantageous. In order for liability to function well as an incentive device, injurers must have assets approximating the harm they might cause, whereas regulation and taxation (based on expected harm rather than actual harm) do not require injurers to have substantial assets. Liability, however, may enjoy an administrative cost advantage over regulation and taxation, in that administrative costs are incurred under the liability system only when harm comes about, whereas such costs generally are incurred more often under regulation and taxation. On the comparison of the liability system and other means of controlling risk, see Calabresi and Melamed (1972), Kolstad et al. (1990), and Shavell (1993).

See Also

- ▶ Externalities
- ▶ Law, Economic Analysis Of

Bibliography

Brown, J.P. 1973. Toward an economic theory of liability. *Journal of Legal Studies* 2: 323–349.

Calabresi, G. 1970. *The costs of accidents*. New Haven: Yale University Press.

Calabresi, G. 1975. Concerning cause and the law of torts. *University of Chicago Law Review* 43: 69–108.

Calabresi, G., and A.D. Melamed. 1972. Property rules, liability rules, and inalienability: One view of the cathedral. *Harvard Law Review* 85: 1089–1128.

Cooter, R.D. 1989. Punitive damages for deterrence: When and how much? *Alabama Law Review* 40: 1143–1196.
Craswell, R., and J.E. Calfee. 1986. Deterrence and uncertain legal standards. *Journal of Law, Economics, and Organization* 2: 279–303.
Diamond, P.A. 1974. Single activity accidents. *Journal of Legal Studies* 3: 107–164.
Kahan, M. 1989. Causation and incentives to take care under the negligence rule. *Journal of Legal Studies* 18: 427–447.
Kessler, D., and D.R. Rubinfeld. 2007. Empirical study of the common law and legal process. In *Handbook of law and economics*, vol. 1, ed. A.M. Polinsky and S. Shavell. Amsterdam: North-Holland.
Kolstad, C.D., T.S. Ulen, and G.V. Johnson. 1990. Ex post liability vs. ex ante regulation: Substitutes or complements? *American Economic Review* 80: 888–901.
Kornhauser, L. 1982. An economic analysis of the choice between enterprise and personal liability for accidents. *California Law Review* 70: 1345–1392.
Landes, W.M., and R.A. Posner. 1987. *The economic structure of tort law*. Cambridge, MA: Harvard University Press.
Pitchford, R. 1995. How liable should a lender be? The case of judgment-proof firm and environmental risk. *American Economic Review* 85: 1171–1186.
Polinsky, A.M., and D.R. Rubinfeld. 1988. The welfare implications of costly litigation for the level of liability. *Journal of Legal Studies* 17: 151–164.
Polinsky, A.M., and S. Shavell. 1998. Punitive damages: An economic analysis. *Harvard Law Review* 111: 869–962.
Priest, G.L. 1981. A theory of the consumer warranty. *Yale Law Journal* 90: 1297–1352.
Rubin, P.H. 1993. *Tort reform by contract*. Washington, DC: AEI Press.
Shavell, S. 1980. Strict liability versus negligence. *Journal of Legal Studies* 9: 1–25.
Shavell, S. 1982a. On liability and insurance. *Bell Journal of Economics* 13: 120–132.
Shavell, S. 1982b. The social versus the private incentive to bring suit in a costly legal system. *Journal of Legal Studies* 11: 333–339.
Shavell, S. 1986. The judgment proof problem. *International Review of Law and Economics* 6: 45–58.
Shavell, S. 1987. *Economic analysis of accident law*. Cambridge, MA: Harvard University Press.
Shavell, S. 1993. The optimal structure of law enforcement. *Journal of Law and Economics* 36: 255–287.
Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *Journal of Legal Studies* 26: 575–612.
Shavell, S. 2004. *Foundations of economic analysis of law*. Cambridge, MA: Harvard University Press.
Shavell, S. 2005. Minimum asset requirements and compulsory liability insurance as solutions to the judgment-proof problem. *Rand Journal of Economics* 36: 63–77.
Spence, M. 1977. Consumer misperceptions, product failure, and producer liability. *Review of Economic Studies* 44: 561–572.
Sykes, A. 1984. The economics of vicarious liability. *Yale Law Journal* 93: 1231–1280.

Liberalism and Economics

Ralf Dahrendorf

Keywords

Anarchism; Beveridge, W. H.; Buchanan, J. M.; Capitalism; Classical liberalism; Conservatism; Democracy; Friedman, M.; Hayek, F. A. von; Invisible hand; Keynes, J. M.; Liberalism; Liberty; Locke, J.; Market; Minimal state; Natural rights; Neoliberalism; Nozick, R.; Popper, K.; Rule of law; Smith, A.; Social contract; Social democracy; Social liberalism; Social market economy; Social rights; Socialism; Stagflation; supply-side economics; Totalitarianism; Uncertainty; Unemployment; Welfare state

JEL Classifications

B0

Liberalism is the theory and practice of reforms which has inspired two centuries of modern history. It grew out of the English Revolutions of the 17th century, spread to many countries in the wake of the American and French Revolutions of the 18th century, and dominated the better part of the 19th century. At that time, it also underwent changes. Some say it died, or gave way to socialism, or allowed itself to be perverted by socialist ideas; others regard the social reforms of the late 19th and 20th centuries as achievements of a new liberalism. More recently, interest in the original ideas of liberals has been revived. Thus, classical liberals, social liberals and neoliberals may be distinguished.

Classical liberalism is a simple, dramatic philosophy. Its central idea is liberty under the law.

People must be allowed to follow their own interests and desires, constrained only by rules which prevent their encroachment on the liberty of others. Early liberals before and after John Locke (1690) liked to use the metaphor of a social contract to express this view. Society can be thought of as emerging from an agreement among its members to protect themselves against the selfish desires of others. Man's 'unsociable sociability' (Kant 1784) makes rules necessary which bind all, but requires also the maximum feasible space for competition and conflict.

In fact, of course, early liberals were not concerned with building societies from scratch. They were concerned with forcing absolute rulers to yield to demands for liberty. The rule of law envisaged by liberals was a revolutionary force which heralded the enlightened phase of modernity.

The notion, rule of law, is not without ambiguity. It is, in the first instance, largely formal. One thinks of rules of the game applying to all and regulating the social, economic and political process. In theory, such rules are intended not to prejudge the outcome of the game itself. Still, even their formal conditions, equality before the law and due process, involved fundamental changes which justify speaking of a movement of reform. Throughout the history of liberalism, however, the question of certain substantive rights of man has been an issue. The inviolability of the person and the rights of free expression have been liberal causes along with constitutional rules. Liberals have rarely found it easy to reason for such substantive rights to their own satisfaction. A certain tension between liberal thought and the notion of natural rights is unmistakable.

The modern debate of these issues began in Scotland and England. John Locke, David Hume (1740) and Adam Smith (1776) are but three of many names to consider. From Britain, the ideas spread to the United States and to continental Europe. Montesqieu and Kant borrowed some of their ideas from British liberals. The American Declaration of Independence and the Constitution, the Declaration of the Rights of Man three years after the French Revolution are only two practical illustrations of the effect of the new ideas. If one wants to, one can distinguish, with Friedrich von Hayek, between a British 'evolutionary' and a continental 'constructivist' concept of liberalism. Either or both however became the dominant reform movements of the early 19th century and determined the dynamics of Europe and North America between the 1780s and the 1840s or 1850s.

Liberalism had consequences for economic, social and political thought. Its economic application was the most obvious and remains the most familiar. If rules of the game are all that can be justified whereas otherwise interests should be allowed a free reign, the scene is set for the operation of the market. It is the forum where equal rights of access and participation but divergent and competing interests lead, through the operation of an 'invisible hand' (Adam Smith), to the greatest welfare for all. Liberalism and market capitalism are inseparable, much as later European theorists (notably in Germany and Italy) have tried to dissociate the two.

The social application of liberalism analogously leads to the emergence of the public, if by 'public' we understand the meeting place of divergent views from which a 'public opinion' emerges. On the Continent, a more emphatic language is often preferred; here, one likes to speak of the emergence of society from under the state. Either way, the basic idea involves the same departure from an all-embracing system of domination by traditional authorities to one in which public authority is confined to certain tasks of regulation, and thus bound to grant and defend the freedom of individuals to express their views.

This is the point at which classical liberalism was not only instrumental for the promotion of market capitalism and social participation, but also for the development of what is called today, democracy. Again, the term is anything but clear. It can be understood to mean a system of government which is based on the competition of divergent views – individual views or group views – for power, constrained by rules which limit the instruments used in the process, and stipulate the possibility for change. In this sense, a variety of constitutional forms of democracy respond to liberal views, including versions of representative

government as well as forms of plebiscite. Liberalism is not anarchism, but anarchism is in some ways an extreme form of liberalism. The law has a key role in liberal thinking, but for a long time the prevalent interest of liberals was that of liberating people from the fetters of control imposed by the tangible force of the state (and the church) or the abstract force of tradition. Not surprisingly, some authors took this intention of liberation to its extreme. If they believed in the essential goodness of man, they advocated the abolition of all social restraint; at times, Jean-Jacques Rousseau seems to argue this way. If on the other hand they believed in the ambivalence of human nature, they were not afraid to demand unlimited room for manoeuvre for 'the singular one and his property' (Max Stirner 1845).

Perhaps this anarchist strain in early liberal thinking can be said to have been one of the reasons for the counter-reaction of the 19th century. Marx was the first to point out the historical advance brought about by 'bourgeois' equality before the law, including the contractual basis of economic action, but also the price paid by many for the 'anarchic' quality of the resulting market. The market – it was increasingly argued – was in fact not neutral, but favoured certain players to the systematic disadvantage of others. Mass poverty, conditions of labour, the state of industrial cities were cited as examples. Nor was this merely a view of anti-liberals. The great ambiguities in the thinking of John Stuart Mill tell the story.

There are two ways of describing the resulting history of thought and of social movements. One is to say that as the 19th century progressed, and certainly in the early decades of the 20th century, liberalism was replaced by socialism as a dominant force. People began to shrink back from the unconstrained market and sought new kinds of intervention. Today, authors would add that the 'structural change of the public' (J. Habermas 1962) and the bureaucratization of democracy followed suit. Liberalism died a 'strange death'; it ceased to be a source of reform and became a defence of class interest.

Another view ascribes the new reforms to liberals also, albeit to a different kind of liberalism. In his Alfred Marshall Lectures of 1949, T.H. Marshall (1950) argued that the progress of citizenship rights had to involve, from a certain point onwards, their extension from the legal and the political to the social realm. Social citizenship rights turned out to be a necessary prerequisite for the exercise of equality before the law and universal suffrage. Thus, the social, or welfare state was no more than a logical extension of the process which began with the revolutions of the 18th century.

There is much to be said for this line of argument if one considers that the two men who above all determined the climate of political thought and action from the 1930s to the 1970s, John Maynard Keynes and William Beveridge, were both self-declared liberals. In effect if not in intention, they advanced ideas which led to restrictions on the operation of markets. One will be remembered as the author of economic policy as a deliberate effort by governments, the other has contributed much to the creation of transfer systems which are operated by governments in the light of an assumed common interest. In other words, these were liberals who pursued policies which led to strengthening rather than limiting the power of public authorities. Theirs was a substantive, a social liberalism.

Liberal parties have found it difficult to follow the twists of theoretical liberalism. Before the First World War, when socialist parties were still in their infancy and unable to determine policy in any major country, they were often the spokesmen of the deprived and underprivileged. At least one strand of the liberal tradition continued to be reformist. However, after the First World War, socialists or social democrats came to form governments in many countries. Their gain was the liberals' loss. Liberal parties declined to the point of insignificance, unless they merely kept the name and changed their policies out of recognition, either in the direction of social democracy (Canada) or in that of conservatism (Australia). Indeed, as a practical political movement, liberalism came to present such a confused picture that Hayek could argue that liberalism has become a mere intellectual, and not a political force.

The experience of totalitarianism interrupted this process without stopping it altogether. To

the dismay but also to the surprise of many, basic human rights and the rules of the game of civil government became an issue again in the 1930s and 1940s. This gave rise to an important literature in which the underlying values of liberal thought were spelt out anew. Hayek's *Road to Serfdom* is one example, but the most important one is probably Karl Popper's *Open Society and Its Enemies* (1952). Popper developed above all what might be called the epistemology of liberalism. We are living in a world of uncertainty. Since no one can know all answers, let alone what the right answers are, it is of cardinal importance to make sure that different answers can be given at any one time, and especially over time. The path of politics, like that of knowledge, must be one of trial and error. The principle can be applied to economy and society as well.

The liberal revolt against totalitarianism waned with the memory of totalitarianism itself. While the term 'social market economy' was coined for Germany in the 1950s, the quarter-century of the economic miracle was in fact a social-democratic quarter-century. In it, economic growth was combined almost everywhere with a growing role of government and with the extension of the social state. Entitlements came to matter as much as achievements. Consensus counted for more than competition or conflict. Despite variations, this was a very successful period in the countries of the First World. But by the 1970s, the side effects of success had become major problems in their own right. These were not only obvious problems like environmental and social 'limits to growth', but systematic ones arising from the role of the state. Both Keynes and Beveridge gave rise to new questions. Neither stagflation in the 1970s nor boom unemployment in the 1980s seemed amenable to government intervention. The social state had got out of hand; it became harder and harder to finance, and its bureaucracies robbed it of much of its plausibility. There were demands for a reversal of trends.

Where such a reversal happened, it remained bitty, halting and inconsistent. However, the new climate gave rise also to elements of a new theory of liberalism. In one sense, this was, and is a return to the original project of asserting society against the state, the market against planning and regulation, the right of the individual against overpowering authorities and collectivities. American authors in particular restated the theory. Milton Friedman tried to show in a series of arguments that the role of government is usually contrary to the interests of people. Robert Nozick made a strong case for the 'minimal state' and against the arrogance of modern state power. James Buchanan (1975) and the 'constitutional economists' reconstructed the social contract and argued for severely limited rules and regulations, using the fiscal system as one of their main examples. This trend, more than the notion of supply-side economics (which in some ways is merely Keynes stood on his head) signifies the revival of liberalism.

There are other facets of the many-faceted term. For many, the extension of civil rights to hitherto disadvantaged groups is a liberal programme. Others still concentrate on the separation of church and state and the reduction of church influence. Again others regard liberalism as an advocacy of cultural values, including pluralism and creativity. It is not difficult to see the connection of such preferences with the mainstream of liberal thought.

This mainstream has three elements. Liberalism is a theory and a movement of *reform to advance individual liberties* in the horizon of *uncertainty*. This means by the same token that the prevailing theme of liberalism cannot be the same at all times. In the face of absolutism, it is liberty under the law; in the face of market capitalism, it is the full realization of citizenship rights; in the face of the 'cage of bondage' (Max Weber 1922) of modern bureaucratic government, it is the optimal, if not the minimal state. The struggle for the social contract has become virulent in the advanced free societies. The crisis of the social state, the new unemployment, issues of law and order all raise basic questions of what is Caesar's and what are therefore the proper limits of individual desires. It is no accident that constitutional questions have come to the fore in several countries. At such a time, liberalism is gaining new momentum. It will not solve all issues, but it will remain a source of dynamism and progress towards more life chances for more people.

See Also

- ▶ Invisible Hand
- ▶ Libertarianism
- ▶ Property Rights
- ▶ Utilitarianism and Economic Theory

Bibliography

Buchanan, J. 1975. *The limits of liberty*. Chicago: University of Chicago Press.
Habermas, J. 1962. *Strukturwandel der Öffentlichkeit*. Neuwied: Luchterhand.
Hume, D. 1740. *A treatise of human nature* (Ed. L.A. Selby-Bigge). Oxford: Clarendon Press, 1888.
Kant, I. 1784. Idee zu einer allgemeinen Geschichte inweltbürgerlicher Absicht. In *Kants Populäre Schriften*, ed. P. Menzer. Berlin: Georg Reimer, 1911.
Locke, J. 1690. *Second treatise of government* (Ed. T.P. Peardon). New York: Liberal Arts Press, 1952.
Marshall, T.H. 1950. *Citizenship and social class*. Cambridge: Cambridge University Press.
Popper, K.R. 1952. *The open society and its enemies*, 2nd ed. London: Routledge/Kegan Paul.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford University Press, 1976.
Stirner, M. 1845. *Der Einzige und sein Eigentum*. Leipzig: D. Wigand.
Weber, M. 1922. *Wirtschaft und Gesellschaft*, 4th ed. Töbingen: Mohr/Siebeck, 1956.

Libertarianism

David D. Friedman

Abstract

Libertarians favour coordination by voluntary decentralized mechanisms such as private property and trade. In response to economic arguments for government intervention in the market, they point to the existence in the real world of private solutions to many problems of market failure and the ubiquity of market failure in political markets. Libertarians differ among themselves in the degree to which they rely on rights-based or consequentialist arguments and on how far they take their conclusions, ranging from classical liberals, who wish only to drastically reduce government, to anarcho-capitalists who would replace all useful government functions with private alternatives.

Keywords

Anarchism; Anarcho-capitalism; Antitrust; Arrow, K.; Bork, R.; Coase, R.; Consequentialism; Democracy; Depletable resources; Efficiency; Externalities; First efficiency theorem; Flat tax; Food and Drug Administration (USA); Free trade; Friedman, M.; George, H.; Harsanyi, J.; Higher education; Hotelling, H.; Human capital; Immigration; Imperfect competition; Intellectual property; Land tax; Laissez-faire; Left libertarianism; Libertarianism; Liberty; Limited liability; Lobbying; Locke, J.; Market failure; Negative income tax; Nozick, R.; Objectivism; Paternalism; Positive and negative rights; Pressure groups; Professional licensing; Property rights; Public choice; Public enforcement of law; Public goods; Rand, A.; Rational ignorance; Redistribution of income and wealth; Rent seeking; Rights; Rothbard, M.; Schooling; Status; Tariffs; Tiebout, C.; Utilitarianism; Victimless crimes; Welfare state

JEL Classifications

B1

Libertarians, in current American usage and in this essay, are those who prefer to organize the world through the decentralized mechanisms of private property, trade, and voluntary cooperation rather than through government. Their position is thus a modern variant of the liberalism of the 19th century. Libertarians are likely to be critical of eminent domain, government regulation of business, paternalistic social policies, income redistribution, laws banning 'victimless crimes' such as drug use, gambling and prostitution, and much else. Since there are good arguments for government as well as good arguments against, only a minority of libertarians carry their position all the

way to anarchism. Most accept some level of taxation to pay for the production of public goods such as national defence. Some accept government production or subsidy of things well short of pure public goods, such as schooling.

The term 'libertarian' is also sometimes applied to left anarchists, usually outside of the United States; its original meaning seems to have been believers in free will. The current American usage is largely a response to the shift in the meaning of 'liberal' over the first half of the 20th century. Since believers in what used to be called liberalism could no longer use that term without confusion, many adopted 'libertarian' as a substitute.

One reason for libertarians to support a less than perfectly libertarian society is the belief that, in terms of individual liberty, it is the best we can do. A second is the belief that, while liberty is important, it is not the only thing that is important. Support by many libertarians for government funding of some public goods – scientific research and public health are examples – is based on the idea not that their production makes us freer but that it makes us better off in other ways.

In this article I sketch the general arguments for a libertarian position, discuss libertarian views on particular issues, and finally consider different forms of libertarianism and the internal disagreements that define them.

Why Liberty Is Right

Libertarian conclusions may be supported either by showing that restraints on individual liberty are wrong or by showing that they lead to undesirable consequences. The former approach is often put in terms of individual rights. Each person has a right to control his own body, a right violated by laws against using drugs, by a military draft, and by many other government acts. Each person has a right to control his legitimately acquired property, a right violated by taxation, regulation, price controls,. . ..

Putting the argument in this form raises an obvious question: how to justify such claims. Libertarians offer a variety of answers, ranging from Objectivists, who believe that individual rights can be logically deduced from the nature of man, to intuitionists, who induce them by trying to generalize their moral intuitions (Rand 1964; Den Uyl and Rasmussen 1991; Rothbard 1978; Lester 2000; Nozick 1974; Boaz 1997, 1998).

It also raises questions about how rights are acquired and how far they extend. Almost nobody argues that my right to control my body includes the right to punch you in the nose. Whether it includes the right to make noise on my property that keeps you awake or burn coal in my fireplace whose smoke makes you cough is less clear.

Robert Bork, in the article (Bork 1971) explaining why he was not a libertarian, argued that my disutility from knowing that you are doing something I disapprove of is just as real an externality as my disutility from breathing your smoke, hence that there is no rights-based case for individual freedom as libertarians understand it. If we treat everything I do that affects others without their consent as a trespass liable to be enjoined, we are left with no self-regarding actions and no liberty – the exception swallows the rule. A response from the standpoint of moral philosophy depends on some way of deriving rights that distinguishes between those sources of disutility to me that do and those that do not violate my rights – hitting me over the head versus living your life in a way I disapprove of.

The economic response starts by observing that the enforcement cost of a rule giving me control over my own body is low, since I already control my body. The enforcement cost of giving you control over my body is substantial. Hence the latter alternative is an inefficient definition of property rights, at least unless my use of my body clearly imposes substantial and measurable costs on you that cannot be dealt with by voluntary transactions along Coasean lines. Although your disutility from knowing that I am reading pornography may be just as real as your disutility from breathing my smoke, it is considerably harder to demonstrate to a court, so a liability rule awarding you damages for the disutility you suffer from my reading pornography is likely to result in inefficient outcomes and substantial litigation costs.

Alternatively, a property rule giving you rather than me a property right in my behaviour – requiring me, before doing anything, to get permission from everyone who objects – imposes transaction costs due to the hold-out problem sufficient to guarantee that nobody ever does anything, which is unlikely to be the efficient outcome. Following out this line of argument provides a defence of libertarian conclusions on consequentialist grounds.

'Liberty' and 'rights' are rhetorically powerful words, so it is not surprising that libertarians are not the only ones who claim them. Competing uses can be clarified by distinguishing between negative rights ('the area within which a man can act unobstructed by others', Berlin 1969, p. 122) and positive rights. A negative right is a right to be left alone. A positive right is the right to some outcome. The right not to be killed is a negative right, the right to live – implying the right to be provided with what you need to live, such as food – a positive right. Other positive rights sometimes claimed include the right to decent housing, adequate food, medical care and equal treatment.

One problem with positive rights is that they contradict negative rights, including some that many find persuasive. If I have the right to decent housing and medical care, someone else must have the obligation to produce them, which is inconsistent with his right to control his own body. If I have the right to equal treatment, the right not to have an employer or homeowner decide whether to deal with me on the basis of my race or religion, someone else does not have the right of freedom of association, since he is required to deal with me even if he prefers not to. If I have the right not to be hated or despised for my sexual preferences, that means that I have a claim over the inside of your head, that being where your emotions are to be found. Thus the assertion of positive rights can be seen, and by libertarians often is seen, as the claim that some people are to some degree the slaves of others, required to serve them without having consented to do so – the violation of a deeply held negative right.

A second problem with positive rights is that they are more prone to internal inconsistency than negative rights. There is no conflict between my not killing or enslaving you and your not killing or enslaving me. But there is a conflict between my having adequate food, housing and medical care and your having them, if one or another of those goods happens to be in short supply.

Why Liberty Is Useful

Large parts of the consequentialist argument for individual freedom go back to Adam Smith and should be familiar to every economist. Private property, exchange, prices provide a decentralized coordination mechanism that makes it possible for individuals with different objectives, knowledge and abilities to cooperate while pursuing their separate ends. In the limiting case of perfect competition, the result is provably efficient in the usual economic sense – cannot be improved by even a perfectly intelligent central planner with unlimited control over the actions of the planned. (For both the classical and modern versions of the First Efficiency Theorem, see Arrow 1983, and references therein. For a non-technical sketch of the classical version, see Friedman 1997, ch. 16.)

The fact that this argument is correct, non-obvious, and included in the professional training of any economist is part of the reason why libertarianism is more popular with economists than with most other academics and why even non-libertarian economists tend to be sympathetic to market approaches. To put it differently, one important reason for the rejection of libertarian conclusions by non-economists is the failure to understand price theory – how markets solve the coordination problem.

The Case Against

Yet not all economists, not even all good economists, are libertarians. The economic counter-argument starts with the facts that real markets are imperfectly competitive and real individuals are limited by, at least, imperfect information, transaction costs, and limited calculating ability.

Once we drop the assumptions of the ideal model we are faced with the possibility of market failure, situations where individual rationality fails to lead to group rationality and hence where it is possible for restrictions on the actions of each to produce a better outcome for all. Familiar examples include the underproduction of public goods, the overproduction of negative externalities, and potentially beneficial transactions blocked by adverse selection.

These are real problems, but not always insoluble ones. A market failure results in an outcome inferior, for all concerned, to some alternative outcome. A sufficiently ingenious entrepreneur may be able to create that alternative and collect a share of the net benefit as his reward; a market failure is also a profit opportunity. Radio broadcasts are a pure public good produced privately. So are the services that Google provides to its users. Other forms of market failure may be dealt with by the development of systems of private norms (Ellickson 1991; Posner 2000). Where market failure exists we can expect private arrangements to produce imperfect outcomes, but less imperfect than casual consideration might suggest. (For an interesting example of a real world solution to a theoretically intractable market failure, see Cheung 1973.)

A second objection to the argument for laissez-faire is that efficiency as defined in economics in the sense of Marshall or Hicks–Kaldor (Friedman 1997, ch. 15) is inadequate as a normative criterion, so that a less efficient outcome may be preferable to a more efficient one. What is maximized by the market is value defined by willingness to pay, measured in dollars not utiles, so a transfer from rich to poor might decrease value measured in dollars but increase total utility.

This utilitarian argument for redistribution can be seen as a special case of the argument from market failure. Declining marginal utility is not merely a conjecture of philosophers; it is observed, in the form of risk aversion, in individual choices under uncertainty. In a perfect market, individuals would buy insurance against the risk of being born poor up to the point where the marginal utility costs of any resulting disincentives or transactions costs just balanced the marginal utility gain of transferring income from states of the world where they were rich to ones where they were poor. Thus the outcome of a perfect market would mirror the welfare programme that would be proposed by a utilitarian. It is merely our inconvenient inability to negotiate and sign insurance contracts prior to being born that prevents the market from solving the problem. The argument for utilitarianism in Harsanyi 1955 – that it is what individuals would choose if they were designing a society behind a veil of ignorance with an equal probability of living any of its lives – makes it possible to view redistribution of income either as a way of increasing total utility or as a correction for market failure.

Other objections to market outcomes come from egalitarians who see equality as good in itself and from those who put substantial weight on values unrelated to individual humans achieving their objectives. If what really matters is the preservation of endangered species, whether or not of any value to human beings, there is no guarantee that the market to achieve it. The same is true if what really matters is behaving according to God's will, producing great art and literature, or doing justice whatever the consequences.

A Libertarian Response

It follows that one can imagine outcomes that improve, in one sense or another, on the outcome of pure laissez-faire. It does not follow that one can construct institutions that predictably produce such outcomes.

Consider the case of market failure. It exists because actions taken by A sometimes have effects on B. If A is free to ignore those effects he may make the pair on net worse off by taking actions that increase his welfare by less than they decrease B's or failing to take actions that would increase B's welfare by more than they decrease A's. A well-designed legal structure can sometimes make it in A's interest to take account of those effects, whether through property rules, liability rules, or bargaining between the parties. But sometimes, for reasons explored by Coase (1960) and others (Friedman 2000, pp. 39–45), no legal

structure can be constructed that makes it in the interest of all parties to make the efficient choices.

All this is true in private markets. But it is true far more often in the political markets that control the political institutions that are proposed as a solution to market failure in private markets.

Consider the naive model of democracy – politicians doing good because if they do not they will lose the next election. In order for it to work, individual voters have to acquire the information needed to know what politicians are doing and whether it is good. No politician campaigns on the slogan 'I'm the bad guy'. No farm bill is labelled 'An act to make farmers richer and city folk poorer'.

If I correctly identify the better candidate, vote for him, and – improbably – my vote proves decisive, the benefit is shared with everyone in the polity. The cost is borne by me alone. Time and energy spent acquiring the information necessary for informed voting produce something very close to a pure public good. Public goods are underproduced; one with a public of many millions is likely to be very badly underproduced. The implication is rational ignorance, voters failing to acquire the information they need to judge politicians because its value to them is less than its cost. That eliminates the simple argument for why politicians will find it in their political interest to act as we would wish them to.

A similar problem arises with a more sophisticated model in which political outcomes are driven by interest group pressure. The more an interest group stands to gain by passing or blocking a piece of legislation, the more it will offer politicians in order to support or oppose it. If that were the only relevant factor, the market for legislation would produce something close to an efficient outcome. If a bill produced net benefits, its supporters would spend more supporting it than its opponents spent to block it, and the bill would be likely to pass.

It is not the only relevant factor. An interest group lobbying for legislation is producing a public good for its members and faces an internal public good problem in doing so, since members that refuse to contribute will still benefit if the bill passes. Some interest groups are much better able than others to solve their internal public good problem. A concentrated interest group such as the auto industry – a handful of firms and one union – can raise a substantial fraction of the benefit it expects from an auto tariff in order to lobby for it. A dispersed interest group such as consumers of automobiles and producers of export goods, the people that bear most of the burden of such a tariff, can raise a negligible fraction of the cost to lobby against. Hence we would expect the political market to consistently redistribute from dispersed interest groups to concentrated ones, even when the benefit to the latter is much smaller than the cost to the former – as demonstrated by the continued existence of tariffs nearly two centuries after Ricardo demonstrated that they are, under most circumstances, injurious to the nation that imposes them.

In a private market, a producer receives a price that measures the value to consumers of what he produces, pays a cost that measures the cost to the suppliers of his inputs of producing them, and pockets the difference. It is only when special circumstances arise – externalities that cannot be dealt with by the market, information asymmetry, and the like – that his actions impose net costs or benefits on others. In the political market, in contrast, almost all decisions are made by people who bear few of the costs and receive few of the benefits those decisions produce. A legislator who passes an auto tariff imposes net costs of many billions of dollars on those affected, but all that comes out of his pocket is the extra cost of the car he buys. A judge whose precedent establishes a seriously inefficient legal rule might reduce national income by, say, a tenth of a percentage point – a staggering amount of damage for a single human being to do. But not only will he not pay any of the cost, he will never even know he made a mistake.

Consider, for example, *Davis v. Wyeth Laboratories, Inc.,* 399 F.2d 121 (9th Cir. (Idaho) Jan 22, 1968), where the court found Wyeth liable for the failure to adequately warn of the risk of polio vaccination. Their argument hinged on whether, if warned, Davis might reasonably have chosen not to be vaccinated. The court wrote: 'Thus appellant's risk of contracting the disease

without immunization was about as great (or small) as his risk of contracting it from the vaccine. Under these circumstances we cannot agree with appellee that the choice to take the vaccine was clear.' They reached this conclusion by comparing the 0.9 in a million chance of getting polio from the vaccination with the 0.9 in a million *annual* rate of adult polio from natural causes. Since vaccination provided protection for many years, possibly a lifetime, the proper comparison was to the risk over many years, not one. The court made a mathematical error of more than an order of magnitude, set a precedent which substantially discouraged the development of new vaccines, caused many, perhaps thousands, of unnecessary deaths, and suffered no penalty for doing so.

Market failure is a real problem. It is a problem in ordinary private markets and a much more severe problem in political markets. That is an argument for shifting decisions, so far as possible, from political to private markets – an argument for, not against, the libertarian position.

A possible response is that decisions should be shifted to public markets only where private markets fail. But some degree of market failure can be alleged for almost any activity. Under legal rules permitting government intervention to correct any alleged market failure, intervention can be expected whenever it is politically profitable.

Libertarians vary in how far they are willing to push the arguments that I have just sketched. Consider the case of national defence, a public good with a very large public. The failure to produce it privately at an adequate level is likely to lead to a drastic reduction in liberty. That is an argument sufficiently strong to convince many, although not all, libertarians to include it in the proper functions of government.

So far I have been dealing with arguments based on market failure, but similar point cans be made with regard to other criticisms of market outcomes. It is true that the market takes account of values only to the extent that individuals do; if nobody cares about the survival of the oldest tree in the world or some threatened species of birds, there is no reason to expect the market to preserve it. But the same is true of the political system. It too is driven by the desires of individuals. It just does a much clumsier job of satisfying them.

Indeed, there are some reasons to expect the market to do a better job of serving 'non-economic' values than the political system. Many are things, not that nobody cares about, but only that most people don't, and the market is generally better at providing for small minorities than the political system. A religion followed by a per cent or two of the population has no difficulty getting the market to produce copies of its scriptures. If it is sufficiently unpopular with the majority, it may have problems getting the government to permit them to be printed. A minority in power might be able to do a better job of diverting resources to serve its values, whether religious or environmental, through the political system than through the market. But shifting decisions to the political system for that reason could be a risky gamble.

Another common criticism, but a mistaken one, is that the market ignores the interest of future generations. Future as well as present demand counts. It is worth planting hardwoods today for harvest a century hence as long as the return is at least as great as from alternative investments. Markets allocate resources over time, as Hotelling (1931) showed, in an economically efficient fashion. If it can be predicted that petroleum will be very valuable a century hence, it is profitable to leave it unpumped now so as to sell it then.

This argument depends on secure property rights. It breaks down if oil saved or a tree planted today is likely to be expropriated tomorrow, making holding it for future use a poor gamble. The alternative to decisions by the market is decisions by political mechanisms. Property rights in the political marketplace are much less secure than those in the private marketplace. A president who accepts costs today for benefits 10 or 20 years in the future can be reasonably confident that neither he nor his party will receive credit for those benefits. A dictator, unlike an entrepreneur, rarely has the opportunity to collect the benefit from investments expected to pay off in the future by transferring his long-term assets to a successor in exchange for immediate payment. Hence we would expect political institutions to be much more inclined to

sacrifice the future to the present than market institutions, a conclusion supported by the evidence of environmental policy in the Soviet Union and Social Security in the United States.

What about income redistribution? Here again, the question is not whether there is an outcome that some would prefer to that produced by the market but whether there are institutions that predictably create such an outcome. The equal distribution of votes gives the poor some advantage on the political marketplace, but it may easily be outweighed by the very unequal distribution of other politically relevant resources. Modern governments are observed to redistribute from rich to poor via welfare, from poor to rich by subsidies for art, music, and – the big one – higher education, paid for mostly by state and local taxes and consumed mostly by people from the upper part of the income distribution. (The median family income of US college freshmen in 2001 was $67,200, compared with a median family income for all households of $42,228 – US Census Bureau 2003, Tables 284 and 683. See Gwartney and Stroup 1986, for a discussion of theory and evidence of the consequences of redistributional policies.) Similarly, farm policy provides a subsidy mostly to wealthy farmers and pays for it mainly by a regressive tax in the form of higher food prices.

A second problem with redistribution is rent seeking. In a polity that redistributes, it is in the interest of nearly everyone to spend resources trying to shift the redistribution in his favour, opposing redistribution from him and promoting redistribution to him (Tullock 1967; Friedman 1973, ch. 38; Krueger 1974). The resulting deadweight cost might easily outweigh any utility gain from redistribution.

Issues

Libertarians differ in how far they are willing to carry their libertarianism. In the following discussion I present libertarian positions and the arguments for them while recognizing that in many cases the libertarian position is not supported by all who consider themselves libertarians.

The Easy Cases

Most of the arguments against price control, wage control, rent control, usury laws, and similar restrictions on the terms of market exchange are familiar to any economist. Many libertarians also argue that such restrictions violate individual rights. If I own my body, it is up to me to decide on what terms I will sell my labour to you. If I own my house, it is up to me to decide what terms I am willing to offer to potential tenants and up to them to decide what terms they are willing to accept. Thus many libertarians would reject not only rent and wage control but also legal restrictions on private discrimination in home sales, employment, and the like. (Nozick 1974, ch. 7, provides an extended discussion and defence of a libertarian view of self-ownership.)

Libertarians taking that position may defend it either in terms of individual rights or by arguing that minorities are worse off in a world where such decisions are controlled by government than in one where they are controlled by private contract. State intervention in the US South during the first half of the 20th century provides an obvious example. A prejudiced majority can do a great deal more harm to the minority it is prejudiced against where decisions are made by the government than where they are made privately.

Free trade is another easy case. If building cars in Detroit costs more than growing grain, putting it on ships, sending them out into the Pacific, and having them come back with Hondas on them, we are better off growing our cars instead of building them. A tariff forces us to use the more expensive technology instead of the less expensive; it protects American auto workers from the competition of American farmers, making Americans on the whole worse off. While economists can construct special circumstances in which a trade restriction might benefit the nation that imposed it, such as infant industries that require temporary protection, the restrictions we observe are not those suggested by such arguments: In the U.S., steel and auto are not infant industries. We observe instead the restrictions predicted by the public choice analysis offered earlier, policies that benefit concentrated interest groups at the expense of dispersed interest groups. (For an explanation of

why tariff protection is particularly likely for declining industries such as steel, see Friedman 1997, p. 294.)

Many libertarians find paternalism another easy case, since it contradicts the idea that each individual owns his own body and is free to make choices regarding it. As a practical matter, paternalistic regulations substitute for each individual's decisions about his own welfare the decisions of someone else. The regulator may have expert information the individual lacks, but he lacks both the individual's specialized knowledge about his own circumstances and the individual's incentive to act in that individual's interest. Thus professional licensing, justified as a paternalistic protection of the consumer, is in practice used by professions to reduce competition and so benefit themselves at the expense of their customers. (The classic discussion is Friedman 1962, ch. 9). Similar arguments apply to laws against victimless crimes – the War on Drugs, laws against prostitution and gambling. Individuals might make the wrong decisions for themselves; others should be free to warn them against doing so. But the final decision ought to be made by each individual for himself.

A familiar example of the dangers of such regulation in the United States is the Food and Drug Administration (FDA). Letting a dangerous drug onto the market ends the regulator's career. Keeping a drug off the market for a few more years can do enormous damage – arguably an excess mortality on the order of a hundred thousand lives in the case of beta-blockers (Gieringer 1985. For a webbed discussion, see FDAReview.org.) But damage that appears only in the mortality statistics is very nearly irrelevant, politically speaking. And the connection between overregulation, higher prices and fewer new drugs is still less visible. (See Peltzman 1973, for a classic examination of the effect of regulation on quality and rate of introduction of new drugs.)

Antitrust

There are legitimate arguments, widely supported by economists, in favour of government intervention against monopolies. Even libertarians are troubled by hypotheticals in which one firm owns the only well in the desert and insists on thirsty travellers giving all they own and indenturing their labour for decades into the future in exchange for a drink. Government regulation of monopoly, however, has its own problems. The regulator needs information he is unlikely to have – cost curves and demand curves – in order to force the firm to follow welfare-maximizing rather than profit-maximizing strategies (Friedman 1997, pp. 238–43). And it is far from clear why a real-world regulator, driven by political rather than altruistic incentives, would attempt to regulate in the public interest rather than letting himself be captured by the regulated industry, a concentrated interest well positioned to reward politicians with money and regulators with future jobs (Stigler 1971). An industry that is imperfectly competitive may be imperfectly efficient, but the situation is not improved by giving firms the opportunity to use government regulation, as the US railroad industry used the Interstate Commerce Commission (ICC), to exclude competitors and restrict competition (Kolko 1977).

Such considerations persuade many libertarians that antitrust, both as a legal doctrine and as a basis for regulation, does more harm than good – that we would be better off putting up with any ills private monopoly may produce, since the cure is likely to be worse than the disease (Friedman 1962, pp. 128–9). Others argue that the state need not prevent monopoly but ought not to support it, and can avoid doing so by refusing to enforce contracts in restraint of trade.

Immigration

The economic arguments for free movement of goods apply to capital and labour as well, implying that immigration produces net benefits for the country that permits it, just as free trade produces net benefits for the country that practises it. Freer immigration also produces what many would consider a desirable redistribution, since its major beneficiaries, the immigrants, are much poorer than those who might be made worse off by their move: workers in the country the immigrants go to, capitalists and landowners in the countries they come from.

This assumes a context of voluntary transactions. Some immigrants may come in order to profit by involuntary transactions, private or political – to commit robbery or collect welfare. And new immigrants, once they become citizens and voters, might use the political mechanism to advantage themselves at the cost of the rest of us. Such arguments help explain why not all libertarians support free immigration – despite empirical evidence that, at least under current circumstances, immigrants pay more in taxes than they collect in benefits (Simon 1989, 1995).

The flip side to the 'immigrant as welfare recipient' argument is that, while the existence of a welfare state makes the desirability of free immigration less clear, free immigration makes it more difficult to maintain a welfare state. Free movement of people imposes limits on the ability of governments to exploit those they rule, similar to the limits that market competition imposes on the ability of firms to take advantage of their customers (Tiebout 1956). For libertarians, that is an additional advantage to freer immigration.

Schooling

The usual argument for government provision or subsidy of schooling is that a democracy requires educated voters and an economy educated workers, hence that money spent educating my children benefits you and your children, hence that leaving education to the free market will result in too little.

The first part of that argument might be true, although it is hard to find evidence to support it. The second is simply bad economics. To the extent that education makes a worker more productive, the additional productivity is reflected in his wages; investing in human capital is no more a public good than investing in physical capital. In both cases the investor may receive less than the full value of his investment due to the distorting effect of taxation – some of my additional productivity goes, not to me, but to the Internal Revenue Service. But subsidizing the investment merely shifts the inefficiency to whoever pays the taxes that fund the subsidy.

There may be indirect externalities to subsidized education – a cure for cancer, say. But not all such externalities are positive. By educating my children I make them better able to use the political system to advantage themselves at the expense of your children. By sending my son to Harvard I give him an opportunity to feel superior to your son, who went to Podunk U. That is a benefit to me and my son, a cost to you and yours, and a negative externality produced by my expenditure on education. As Robert Frank (1986) has persuasively argued, one of the things humans care about and economists ought to take account of is relative status.

This example illustrates a common problem with arguments based on externalities. Those making them usually count only externalities that lead to the conclusion they want – positive if they want to subsidize something, negative if they want to ban it. If an activity produces both positive and negative externalities, as many do, and if we are unable to measure them accurately enough to determine the sign of their sum, we do not know whether we should be encouraging the activity or discouraging – in which case it might be wiser to do neither (Friedman 1971).

Another argument for government involvement in schooling is that, since parents act in their own interest rather than that of their children, they may fail to pay the cost of schooling even when it produces a benefit larger than its cost. But shifting the decision to the political system means shifting it, not to children, but to other adults. Adults routinely make large sacrifices on behalf of their children, much more rarely on behalf of other people's children. So while a parent is not a perfect proxy for his children, he may be the best proxy available – a much better one than either the legislature or the teachers' unions.

Other government activities can be supported, and opposed, with similar arguments. Subsidies for basic research can be defended as producing a public good, rejected on the grounds that enough of the benefits can be privatized to make subsidy unnecessary (Kealey 1997), that government involvement diverts too many smart people into whatever field is currently in fashion, and that it subverts the scientific enterprise by converting the search for truth into a search for grants.

The relevance of public good theory is less clear for police and courts, government activities

traditionally accepted by believers in a minimal government. Law enforcers can choose to pursue criminals who commit crimes against those who have paid for their services and not those who have not; England survived with private thieftakers but without police in the modern sense until well into the 19th century. (Davies 2002; Friedman 1995. Both argue that there is no clear evidence that failure of the traditional system was the reason why it was eventually replaced.) Courts can refuse to settle disputes among those unwilling to pay for the service, and some – both private arbitrators and government courts – do. Many libertarians accept the conventional arguments for state provision of police and courts, paid for by taxation; others do not (Friedman 1973, part 3).

There are a few issues where libertarians disagree among themselves about which side is more libertarian. Intellectual property is one example. Some argue that a book or an invention, as the pure creation of a human mind, deserves strong protection. Others regard all intellectual property as coercive, a restriction on how individuals are permitted to use their own material property. Limited liability for corporations is another such. Many libertarians reject it on the grounds that individuals ought to be liable for their actions. Others see it as a legitimate consequence of freedom of association and contract and observe that, while it is possible for a corporation to impose costs it does not have the resources to compensate for, the same is true for an individual.

Foreign policy provides a particularly divisive example. Opponents of the United States in recent decades have been strikingly unfree societies – Hitler's Germany, Stalin's Russia, Mao's China, Ho Chi Minh's Vietnam – making a policy of overthrowing, or at least containing, them attractive to many libertarians. But such a policy is conducted by a government whose competence and motives libertarians find suspect – and badly done interventionism may well be worse than no interventionism (Friedman 1989, ch. 45). Hence many libertarians favour the non-interventionist policy famously advocated by George Washington – peace and friendship with all, entangling alliances with none.

Libertarian: Yes/No or More/Less

Some libertarians propose a bright line definition of who is a libertarian, often along the lines of 'one who believes in never initiating force against another'. One problem with this is that libertarians do not have an entirely satisfactory account of what determines who owns what – in particular, of how unproduced resources, such as land, become property. Without a clear answer to that question, it is sometimes hard to distinguish the initiation of force from the use of force to defend what you justly own.

A second problem is that the bright line definition, taken literally, eliminates almost everyone, including almost all libertarians. Consider a scenario popularized by the late R.W. Bradford, editor of *Liberty Magazine.* You have carelessly fallen out of a 50th storey window. By good luck, you catch hold of the flagpole of the apartment immediately below you and start trying to climb in the window. The owner of the apartment objects that you are violating his property rights – not only by climbing in his window, but by using his flag pole without his permission. Do you let go and fall to your death? Such arguments suggest that 'libertarian' is more usefully defined as a continuum – more libertarian or less rather than libertarian or not.

An issue which has attracted a good deal of attention within the libertarian movement is whether there ought to be any government at all. One faction, sometimes labeled 'minarchist', supports a government that provides, at least, for courts, police, and national defence. The other – anarchists or anarcho-capitalists – argues that, with suitable institutions, voluntary cooperation in a free market can adequately provide all government services worth providing (Friedman 1989, part 3; Rothbard 1978). The latter position can be defended either on the (rights-based) grounds that all other alternatives involve violations of rights or on the (consequentialist) grounds that, just as the free market does a better job than government of building cars and growing food, it could also do a better job of producing laws and defending rights. While the latter claim seems obviously false to many when they first encounter

it, it has proved sufficiently persuasive to be adopted by a significant minority of those seriously involved with libertarian ideas and libertarian argument. (Liberty Magazine Editors 1999.)

Varieties of Libertarianism

Does 'individuals have the right not to be coerced' mean that one should never initiate coercion or that one should act to minimize coercion? If rights are best protected by a tax-supported system of police and courts, should one support such taxes as a way of minimizing rights violations or oppose them as a violation of rights? (Nozick 1974, pp. 28–35, discusses the distinction between rights as side constraints and a 'utilitarianism of rights' and offers arguments for the former.) One answer makes anarchism something close to a moral imperative, the other decides the anarchist/minarchist issue in terms of how well either alternative works.

It is useful for land to be treated as private property. But how does a claimant get ownership? Locke (1689, ch. 5, section 27) famously argued that he did it by mixing his labour with the land – clearing trees, plowing, removing boulders. But that argument included the proviso that there be as much land and as good available for other claimants, since otherwise the first claimants deprive others of the opportunity to claim land themselves. The value of the land is in part site value and in part value due to human effort; how does the owner get a just claim to the former?

Many libertarians avoid these questions by simply accepting existing titles to land. Others argue that such claims are legitimate only if based on a chain of voluntary transfer back to a legitimate appropriation, whether by Lockean mixing of labour with land or some other mechanism. A few, 'geolibertarians' or, more confusingly, 'left libertarians', reject unqualified private ownership of land entirely, arguing for the land tax of Henry George or something similar (Brody 1983; Friedman 1983; Valentyne and Steiner (2000a, b); George 1879.)

For a final variant on libertarianism, consider someone who accepts both the utilitarian argument for redistribution from rich to poor and libertarian arguments against government intervention in the market. He might favour a laissez-faire society combined with some very simple system of redistribution – say a flat tax used to finance a modest demogrant. (The best-known proposal along these lines is the negative income tax; Friedman 1962, pp. 191–5. A more recent version is Murray 2006.) Making the redistribution simple reduces the opportunity for individuals to spend resources trying to shift it in their favour. Putting all redistribution in one form eliminates arguments for other government interventions defended – often implausibly – as helping the poor. While many, perhaps most, libertarians would be reluctant to consider this a fully libertarian position, it provides a possible compromise for those who accept large parts of the consequentialist argument for libertarian policies while remaining unconvinced by libertarian arguments about rights.

See Also

- ▶ Anti-trust Enforcement
- ▶ Externalities
- ▶ Friedman, Milton (1912–2006)
- ▶ International Migration
- ▶ Laissez-faire, Economists and
- ▶ Public Choice
- ▶ Public Goods
- ▶ Rent Seeking
- ▶ Rothbard, Murray N. (1926–1995)

Bibliography

Arrow, K. 1983. An extension of the basic theorems of classical welfare economics. In *Collected papers of Kenneth J. Arrow, volume 2: General equilibrium*. Cambridge, MA: Belknap.

Berlin, I. 1969. *Four essays on liberty*. Oxford: Oxford University Press.

Boaz, D. 1997. *Libertarianism: A primer*. New York: Free Press.

Boaz, D. 1998. *The libertarian reader: Classic and contemporary writings from Lao Tzu to Milton Friedman*. New York: Free Press.

Bork, R. 1971. Neutral principles and some First Amendment problems. *Indiana Law Journal* 47: 1–35.

L

Brody, B. 1983. Redistribution without egalitarianism. *Social Philosophy and Policy* 1: 71–87.
Cheung, S. 1973. The fable of the bees: An economic investigation. *Journal of Law and Economics* 16: 11–33.
Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.
Davies, S. 2002. The private provision of police during the eighteenth and nineteenth centuries. In *The voluntary city: Choice, community, and civil society*, ed. D. Beito, P. Gordon, and A. Tabarrok. Ann Arbor: University of Michigan Press.
Den Uyl, D., and D. Rasmussen. 1991. *Liberty and nature*. Chicago: Open Court.
Ellickson, R. 1991. *Order without law: How neighbors settle disputes*. Cambridge, MA: Harvard University Press.
FDAReview.org. Theory, evidence and examples of FDA harm. Online. http://www.fdareview.org/harm.shtml. Accessed 2 Aug 2006.
Frank, R. 1986. *Choosing the right pond: Human behavior and the quest for status*. New York: Oxford University Press.
Friedman, M. 1962. *Capitalism and freedom*. Chicago: University of Chicago Press.
Friedman, D. 1971. Laissez-faire in population: The least bad solution. Occasional paper, Population Council. Online. http://www.daviddfriedman.com/Academic/Laissez-Faire_In_Popn/L_F_in_Population.html. Accessed 3 Aug 2006.
Friedman, D. 1973. *The machinery of freedom: Guide to a radical capitalism*, 1989. Chicago: Open Court.
Friedman, D. 1983. Comment on Brody 'redistribution without egalitarianism'. *Social Philosophy and Policy* 1: 88–93.
Friedman, D. 1989. *The machinery of freedom: Guide to a radical capitalism*. Chicago: Open Court.
Friedman, D. 1995. Making sense of English law enforcement in the eighteenth century. *University of Chicago Law School Roundtable* 2(2): 475–505.
Friedman, D. 1997. *Hidden order: The economics of everyday life*. New York: Collins.
Friedman, D. 2000. *Law's order: What economics has to Do with law and why it matters*. Princeton: Princeton University Press.
George, H. 1879. *Progress and poverty*, 2003. New York: Robert Schalkenbach Foundation.
Gieringer, D. 1985. The safety and efficacy of new drug approval. *Cato Journal* 5(1): 177–201.
Gwartney, J., and R. Stroup. 1986. Transfers, equality, and the limits of public policy. *Cato Journal* 6(1): 111–137.
Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 64: 309–321.
Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.
Kealey, T. 1997. *The economic laws of scientific research*. New York: Palgrave Macmillan.
Kolko, G. 1977. *Railroads and regulation, 1877–1916*. Westport: Greenwood.
Krueger, A. 1974. The political economy of the rent seeking society. *American Economic Review* 64: 291–303.
Lester, J. 2000. *Escape from leviathan: Liberty, welfare, and anarchy reconciled*. New York: Palgrave Macmillan.
Liberty Magazine, ed. 1999. The liberty poll. *Liberty Magazine,* February, 11–22.
Locke, J. 1689. *Two treatises of government*, 2nd ed. Cambridge: Cambridge University Press, 1967.
Murray, C. 2006. *In our hands: A plan to replace the welfare state*. Washington, DC: AEI Press.
Nozick, R. 1974. *Anarchy, state and Utopia*. Malden: Blackwell.
Peltzman, S. 1973. An evaluation of consumer protection legislation: The 1962 drug amendments. *Journal of Political Economy* 81: 1049–1091.
Posner, E. 2000. *Law and social norms*. Cambridge, MA: Harvard University Press.
Rand, A. 1964. Man's rights. In *The virtue of selfishness*. New York: Signet.
Rothbard, M. 1978. *For a new liberty: The libertarian manifesto,* rev. ed. Lanham: University Press of America.
Simon, J. 1989. *The economic consequences of immigration*. Cambridge, MA: Blackwell.
Simon, J. 1995. *Immigration: The demographic and economic facts.* Washington, DC: Cato Institute. Online. http://www.cato.org/pubs/policy_report/primmig.html. Accessed 3 Aug 2006.
Stigler, G. 1971. The theory of economic regulation. *Bell Journal of Economics and Management Science* 2: 3–21.
Tiebout, C. 1956. A pure theory of local public expenditures. *Journal of Political Economy* 64: 416–424.
Tullock, G. 1967. The welfare costs of tariffs, monopoly and theft. *Western Economic Journal* 5: 224–232.
US Census Bureau. 2003. *Statistical abstract of the United States.* Online. http://www.census.gov/prod/www/statistical-abstract-2001_2005.html. Accessed 3 Aug 2006.
Valentyne, P., and H. Steiner (eds.). 2000a. *The origins of left libertarianism: An anthology of historical writings*. New York: Palgrave.
Valentyne, P., and H. Steiner (eds.). 2000b. *Left libertarianism and its critics: The contemporary debate*. New York: Palgrave.

Liberty

Alan Ryan

In *The Philosophy of History* Hegel declared that the history of the world was the history of freedom. Human history had been a process of education and self-discovery at the end of which men could see that freedom and reason were their very

essence. But Hegel was conscious of the fact that 'freedom' was not the same thing at all stages of its history – a truth which is reflected in the inconclusiveness of philosophical attempts to answer the question 'what is freedom?'.

The first discussions of freedom in Western political thinking occur among the Greeks. Aristotle, for instance, defines politics as a way of life in which free men rule one another in turn; freedom is a matter of citizenship. The 'unfree' are all those who are subject to an authority to which they have not given their consent. The Persians are ruled despotically and are unfree; women have no role in politics and are dependent on their husbands; slaves are in all things dependent on their owners; and if manual workers are not exactly unfree, they cannot share in the freedom of citizenship because they lack the leisure and intelligence so to do. Freedom is a uniquely Greek possession, for only among the Greeks is politics possible.

Aristotle's discussion did not reflect a Greek consensus; Pericles' 'Funeral Speech' praised Athens's freedom, but used the term in a way any 20th-century reader understand – the Athenians were tolerant of diversity, and did not think that political unity depended on depriving themselves of a vigorous private life. But like Aristotle, Pericles took it for granted that women and slaves were condemned to obscurity and that their freedom raised no questions. Plato was positively hostile to freedom, which he identified with an unbridled opportunity to do whatever we liked; democracy was addicted to liberty – so much so, he said in *The Republic,* that in the streets of Athens even the donkeys will not move out of your way. It was not freedom but discipline based on philosophical illumination which would make individuals good and society stable.

Roman political thinking made three crucial contributions to the discussion. One was a sophistication of the argument about the best way to avoid the tyranny of one man or of a social class; republican freedom was best preserved by a 'mixed constitution' with elements of monarchy, aristocracy and democracy. A second was the Roman Law concern with the status and the nature of slaves; they were archetypically unfree, being as the lawyers said 'legally dead', or 'always children'. They had no legal personality and were always (even when they were rich and powerful administrators) vulnerable to the loss of everything they possessed, since they had no legal protection. The third was the development of an unpolitical conception of freedom, epitomized by the declaration of the Stoic emperor Marcus Aurelius that the slave could be as free as the emperor. A slave who was sufficiently immune to the ills of this life could exercise 'self-control' – and what was freedom if not the condition of autonomy, or controlling oneself?

The rediscovery of both the Greek and the Roman interest in freedom had to wait until the end of the medieval period; but arguments about political liberty did not. Under the often notional overlordship of the Holy Roman Empire, 'free cities' sprang up, and in Northern Italy independent republics never quite severed their links with the ancient tradition. In Florence or Venice, liberty was well understood to be a condition of self-government immune from the interference of foreign powers or from the tyranny of local aristocrats and despots. It is to be noticed that in no discussion of liberty in this sense is there any suggestion that economic freedom or an approach to laissez-faire is part of freedom so understood. Indeed, since the crucial issues were how to secure a reliable militia on the one hand and a non-tyrannical ruling class on the other, the search for public-spirited citizens and rulers generally resulted in the condemnation of self-interest and the lure of wealth.

But an alternative conception of liberty was slowly growing as the feudal and military basis of land ownership was eroded, and something closer to modern commercial relations grew up. Dating the stages of this transformation is difficult, but it is not implausible to talk of the emancipation of English landed property from its feudal past from the 13th century onwards. Allied to this process was an extension of the scope of centralized justice and a sophistication of procedures for protecting legal rights, such that their joint effect was to bring into existence the idea that liberty was above all a matter of being able to enjoy one's rights as a private person, and to know

that these rights set limits to what the holders of power could do and to how they could do it.

There were thus two competing ways of thinking about political freedom waiting to be taken up in the 16th and 17th centuries, while political practice was also affected by Protestant antinomian doctrines, which held that true freedom was not of this world but of the next, a view which might lead to political quietism or to radical outbreaks such as the Anabaptists' revolution in Munster. All could legitimately claim to be theories of liberty; but they naturally led in very different directions. Those who looked back to the ancient republics for their image of liberty thought the triumph of laissez-faire and security for the individual a triumph for 'corruption'. Their opponents thought them nostalgic and unrealistic. The unworldly were attacked as disturbers of the peace when they insisted on following their own consciences rather than their rulers' edicts, condemned as apathetic if they were quietist.

Serious political writers of the 18th century were pulled in different directions. Adam Smith's attitude towards laissez-faire is now generally recognized to be ambivalent; the simple system of natural liberty has its drawbacks, and these drawbacks are explained in 'republican' terms. They include the inability of the common man to display the military valour of his Roman forebears and the danger that wealth will corrupt the institutions which protect us from tyranny. Rousseau anathematized the 18th century's concern with private welfare rather than public spirit and hankered after Spartan simplicity – but agreed that Sparta could not easily be rebuilt in modern Europe. To him we owe a strikingly lucid categorization of liberty: natural liberty is what we enjoy when we are subject to no restraint, moral liberty what we enjoy when we only follow rules which an impartial benevolence would urge upon us, and civil liberty what we enjoy when we are citizens participating in the creation of the laws we obey. It cannot be said that he is wholly successful in explaining which of these we can expect to achieve in the modern European state; he may have hoped that the rule of law would realize moral liberty – his Jacobin readers hoped to recreate Roman civil liberty in a virtuous republic.

The disasters of the Revolution provoked Benjamin Constant to write his *Essai sur la liberté des anciens comparée à celle des modernes,* in which he contrasted the freedom enjoyed by the citizens of the ancient city states with that enjoyed by modern man. Theirs was essentially political, a matter of the right to take part in politics and be a member of the sovereign authority; ours is essentially private, a matter of security under the law and the right to pursue our own goals without interference. Since we cannot recreate the social conditions of the Greek city state, we cannot have ancient liberty. Nor do we really want it, since we want neither the slavery on which liberty for the male citizen population depended nor the continual wars in which the ancient republics engaged. Political rights of the kind enshrined in the American Constitution are essential to preserve modern freedom, but they are needed for self-defence, not to turn us into Roman citizens.

Much subsequent argument has concentrated on the question of how much freedom modern man enjoys in fact. The argument has moved away from political liberty to a wider and less precisely delimited concern with social, economic and psychological constraints. Mill's *Liberty* was dedicated to the proposition that although the English enjoyed political freedom, they did not enjoy social freedom. Public opinion constrained all but the very boldest spirits; the threat of social disapproval made most people afraid to think differently from the majority on any issue whatever. In some ways this tyranny was worse than many forms of political tyranny because it aroused less opposition and worked silently and insidiously. Mill opposed this conformism with two principles – negatively, we must coerce others only in self-defence and not merely for the sake of having them think like us; positively, we must see that individuality is the distinguishing mark of humanity and that we are only free when our ideas and lives are our own. Similar views have been put forward by sociologists and psychologists ever since; many have thought the tyranny of the majority less impressive than our capacity to take away our own liberty by one means or another, but all have held that political liberty,

vital as it is, can only be the beginning of complete liberty.

The claim that under laissez-faire there is less freedom than its defenders suppose goes back to the 18th century. But it has been the central issue between defenders of capitalism and their socialist critics for the past century and a half. Those who identify liberty and 'free enterprise' argue that the man who has no property or no marketable skills is worse off than the man with property or abilities to his name, but no less free. Liberty is counterfactual: *if* he were able to make his way in the world, nobody would prevent him. The poor man cannot dine at the Ritz, but that is not because he is not free to do so. To this is often added a claim which has been popular since Sir Isaiah Berlin's *Two Concepts of Liberty* (1958): freedom is negative, a matter of the absence of coercion. We must not confuse the negative matter of what we are at liberty to do with the positive matter of what we are able to do.

Opponents have seized on different issues. Marx, for instance, claimed that the distinction between rich and poor under capitalism originated in the forcible expropriation of the small freeholder, so that coercion underpinned 'free exchange'. Mill argued that the range of choice open to the poorest manual worker was as limited as that of the slave. Many writers have denied that the distinction between the presence of opportunities and abilities on the one hand and the absence of coercion on the other is as important as has been made out. Others have widened the argument, pointing out that we talk of 'having no choice' where circumstances as well as people dictate what we must do. Following Marx, their strictures on the unfreedom of capitalism sometimes imply a tyranny of capital over all its victims, sometimes a tyranny of capitalists over workers, and sometimes that capital itself forces capitalists to force workers into exploited labour. The attempts of philosophers to show that this is not an argument about liberty – as opposed to justice or happiness or self-respect – have been unsuccessful, but this is far from saying that anyone has a clear idea of just what *would* constitute economic freedom.

See Also

- ▶ Anarchism
- ▶ Individualism
- ▶ Mill, John Stuart (1806–1873)

Bibliography

Berlin, I. 1969. *Four essays on liberty*. Oxford: Oxford University Press.

Hegel, G.W.F. 1822–30. *Lectures on the philosophy of world history: Introduction*. Trans. H.B. Nisbet. Cambridge: Cambridge University Press, 1975.

Lukes, S.M. 1985. *Marxism and morality*. Oxford: Oxford University Press.

Macpherson, C.B. 1973. *Democratic theory*. Oxford: Clarendon Press.

Mill, J.S. 1859. *On liberty*, , 1974. Harmondsworth: Penguin Books.

Nozick, R. 1974. *Anarchy, state and utopia*. New York: Basic Books.

LIBOR: Origins, Economics, Crisis, Scandal and Reform

David Hou and David Skeie

Abstract

The London Interbank Offered Rate (LIBOR) is a widely used indicator of funding conditions in the interbank market. As of 2013, LIBOR underpins more than $300 trillion of financial contracts, including swaps and futures, in addition to trillions more in variable rate mortgage and student loans. LIBOR's erratic behaviour during the financial crisis fuelled market instability, simultaneously provoking questions surrounding its credibility. Ongoing regulatory investigations have uncovered misconduct by a number of financial institutions. Policymakers across the globe now face the task of reforming LIBOR in the aftermath of the scandal and crisis.

Keywords

LIBOR; Financial crisis; Scandal; Interbank; Banking; Reference rate; Interest rate

JEL Classification
G01; G12; G15; G18; G21; E43

Overview

The London Interbank Offered Rate (LIBOR) is the reference rate at which large banks indicate that they can borrow short-term wholesale funds from one another on an unsecured basis in the interbank market. Beginning in 2007, regulators and market observers noted that LIBOR had failed to behave in line with expectations, given other market prices and rates. Commodity Futures Trading Commission (CFTC) investigations in the USA uncovered explicit manipulation by banks to influence rate fixings, with the intent of projecting financial soundness during the crisis and benefiting proprietary trading positions. Three banks – Barclays, UBS and RBS – have combined to pay settlements upward of $2.5 billion. A collaborative effort on the part of policymakers internationally is under way to reform the reference rate.

History and Methodology

LIBOR's origination has been credited to a Greek banker by the name of Minos Zombanakis, who in 1969 arranged an $80 million syndicated loan from Manufacturer's Hanover to the Shah of Iran based on the reported funding costs of a set of reference banks (Ridley and Jones 2012). In addition to providing loans at rates tied to LIBOR, banks whose submissions determined the fixing had also begun to borrow heavily using LIBOR-based contracts by the mid-1980s, creating an incentive to underreport funding costs. As a result, the British Bankers' Association (BBA) took control of the rate in 1986 to formalise the data collection and governance process. In that year, LIBOR fixings were calculated for the US dollar, the British pound and the Japanese yen. Over time, the inclusion of additional currencies and integration of existing ones into the euro left the BBA with oversight over 10 separate fixings as of 2012. Fifteen maturity terms were reported for each currency, ranging from overnight to a 1 year term. However, the number of currency– maturity pairs has fallen in the aftermath of the LIBOR probes (Table 1).

As of October 2013, the BBA is still nominally responsible for administering LIBOR and publishes the rate each business day at approximately 11:30 GMT (06:30 EST). Actual collection of responses and calculations is performed by Thomson Reuters. The official LIBOR fixing for each currency–maturity pair is calculated as the interquartile trimmed mean of submissions: the set of individual bank submissions are ordered, then the top and bottom four responses are discarded, and the remaining values are averaged to arrive at the LIBOR fixing for that currency–maturity pair. The banks that comprise the LIBOR panel are typically the largest and most creditworthy ones with London operations, with the constituents varying based on currency. Of the

LIBOR: Origins, Economics, Crisis, Scandal and Reform, Table 1 Active and inactive LIBOR currencies and maturities as of 12 October 2013

LIBOR Currencies		LIBOR Maturities	
Active	*Inactive*	*Active*	*Inactive*
U.S. Dollar	Australian Dollar	1 Day	2 Weeks
Euro	Canadian Dollar	1 Week	4 Months
British Pound Sterling	New Zealand Dollar	1 Month	5 Months
Japanese Yen	Danish Krone	2 Months	7 Months
Swiss Franc	Swedish Krona	3 Months	8 Months
		6 Months	9 Months
		12 Months	10 Months
			11 Months

10 LIBOR currencies that were reported in recent years, nine had panels consisting of 16 respondents, yielding precisely an interquartile trimmed mean. The USD panel, on the other hand, has 18 respondents as of 12 October 2013, yielding a 23% trimmed mean after the top and bottom four submissions are discarded.

The survey question posed to the panel banks is the following: 'At what rate could you borrow funds, were you to do so by asking for and then accepting interbank offers in a reasonable market size just prior to 11 am?'.

The shortcomings of this survey methodology have come under the spotlight in recent years. Key phrases in the survey question pertaining to timing and size are highly subjective and open to interpretation. A 'reasonable market size' and 'just prior to 11 am' may have different meanings for different respondents, though Ellis (2011) has suggested a few 100 million dollars as the industry standard for the former. Perhaps most importantly, the offer rate being calculated is a hypothetical one not based on actual market transactions. An institution claiming an ability to borrow $100 million for 3 months at 350 basis points (bps) is not required to corroborate that assertion with factual evidence. In theory, the trimmed mean result should correspond closely with actual market transactions, though parity need not necessarily hold in practice.

LIBOR Usage and Substitutes

LIBOR serves two primary purposes in modern markets: as a reference rate and as a benchmark rate. A reference rate is a rate that financial instruments can contract upon to establish the terms of agreement. A benchmark rate reflects a relative performance measure, often for investment returns or funding costs. LIBOR serves as the primary reference rate for short-term floating rate financial contracts like swaps and futures. At its peak, estimates placed the value of such contracts at upwards of $300 trillion (Brousseau et al. 2009; Chen 2013; Ellis 2011; Gensler 2012). (Other sources have estimated values as high as $800 trillion (Wall Street Journal 2013).) Variable rate loans, primarily adjustable rate mortgages (ARMs) and private student loans, are also often tied to LIBOR. As a benchmark rate, it is also an indicator of the health of financial markets. The spreads between LIBOR and other benchmark rates can signal changing tides in the broad financial environment.

The rationale for the wide usage of LIBOR in contracts stems from its construction. Because LIBOR represents the terms at which the world's largest and most financially sound institutions are able to obtain funding on a short-term basis, it serves as the lower bound for the borrowing rate of other less creditworthy institutions and individuals, *ceteris paribus*. Rates are typically expressed as 'LIBOR + x', where x is the premium charged in basis points for each particular borrower on top of the LIBOR rate of the corresponding maturity term. The financial contracts most commonly tied to LIBOR include interest rate swaps and other derivatives, fixed income securities and ARMs. In this sense, banks extending variable rate loans can guarantee a positive net interest margin by ensuring that the interest rates they charge are tied to their cost of funds, with a positive premium built in.

LIBOR's growth to prominence as a reference rate is closely tied to the historical popularity of unsecured term interbank borrowing rates. A Bank for International Settlements (BIS) working group notes that these rates were the first to be introduced and have evolved over time into the industry standard because of early adoption by market participants (BIS 2013). More generally, however, reference rates allow for easier standardisation of financial contracts while reducing the complexity with which terms on floating rate legs are determined. Recent episodes have also underscored the potential weaknesses of a universally adopted reference rate. Adequate market liquidity and depth – a rare concern prior to the financial crisis – has emerged as a top criterion for regulators. Prudent oversight and robustness, even under financial duress, are now necessary components of any conversation about reference rates.

Although the USD LIBOR fixing is the most dominant and widely recognised benchmark rate

in the world, many other reference rates exist that seek to capture funding conditions in global financial markets. EURIBOR is perhaps the second most widely used benchmark rate, next to LIBOR, and is calculated based on the funding abilities of a larger panel of European banks. (Though both rates reflect measures of term borrowing for wholesale euro deposits, EURIBOR is more widely used than LIBOR for the euro currency. Widening spreads between the two rates during the crisis provoked questions of misconduct.) Other financial centres, like Tokyo, Mumbai, Singapore and Hong Kong, feature their own internally calculated rate fixings in TIBOR, MIBOR, SIBOR and HIBOR, respectively. The various rates all employ similar methodologies, though they have on occasion arrived at different fixings. Another strand of unsecured interbank borrowing rates relies on past transactions for quotes. The Euro Overnight Index Average (EONIA) is perhaps the best known in this set and serves as a complement to EURIBOR, since the panel of banks are the same for the two rates.

It is worthwhile to examine the theoretical components of LIBOR to better understand its behaviour during the crisis. LIBOR can be thought of as a combination of term and risk spreads:

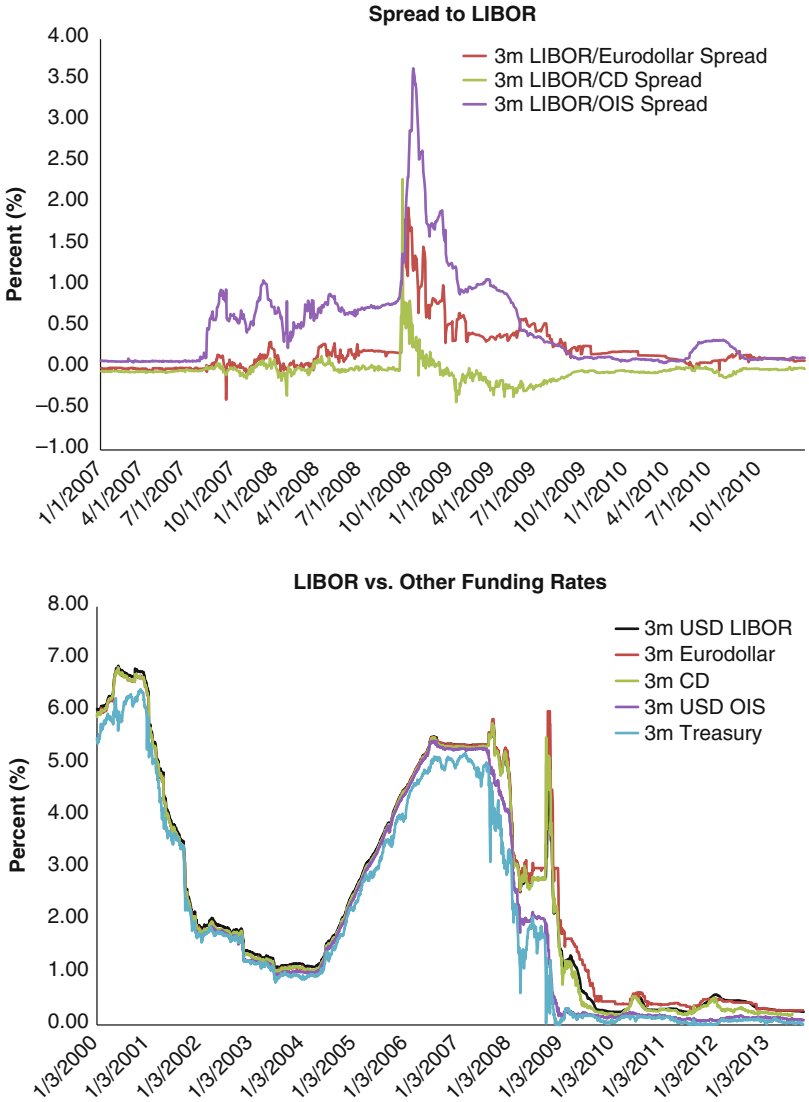
$$\begin{aligned} &LIBOR \\ &= overnight\,risk\,free\,rate\,over\,the\,term \\ &+term\,premium \quad + bank\,term\,credit\,risk \\ &+term\,liquidity\,risk + term\,risk\,premium \end{aligned}$$

The first term is the traditional hypothetical overnight interest rate at which a riskless institution could expect to borrow over the LIBOR loan period. The term premium represents the intertemporal rate of substitution for the term of the loan. Because LIBOR banks are not inherently risk-free borrowers, we must add on the borrower's counterparty *credit risk* component, commensurate with loan maturity. The *term liquidity risk* compensates for maturity risk incurred by the lender by tying up funds for a longer period of time, which could include market illiquidity for interbank funds that may increase the lender's rollover refinancing costs. Finally, the *term risk premium* builds in compensation for the risk that any of these components may have realisations that differ from their expected amounts.

Academic studies have attempted to pin down the fractional contribution to LIBOR attributable to each of these constituent pieces. Acharya and Skeie (2011) attribute the majority of the risk to liquidity, suggesting that liquidity hoarding during stress drives rising interbank rates. This view is shared by McAndrews et al. (2008), Michaud and Upper (2008) and Schwarz (2010), among others. Taylor and Williams (2008a, b), on the other hand, argue that counterparty credit risk as proxied by CDS spreads was the key determinant of driver of interbank rates. Smith (2012) finds that up to 50% of the variation in money market spreads can be explained by the term risk premium.

Behavior During the Crisis

Prior to mid-2007, LIBOR tended to move closely with other short-term interest rates, such as Treasury yields and the Overnight Index Swap (OIS) rate. However, LIBOR began to display erratic behaviour in August 2007 with the onset of the financial crisis. A combination of counterparty credit and liquidity concerns drove the 3-month USD LIBOR to 5.62% on 31 August 2007, compared to an average of 5.36% in the 6 months prior, during a time of stable expectations for the overnight federal funds policy target rate for the Federal Reserve. The maturity-matched OIS rate measures expectations over the tenor of unsecured overnight bank borrowing rates, which in the USA correspond to the effective average federal funds rate. The LIBOR-OIS spread is a measure of the bank credit spread, term liquidity spread and term risk premia for interbank loans (Thornton 2009). This spread is a closely monitored barometer of the health of the banking system and averaged less than 10 bps from 2005 to mid-2007. However, it climbed to more than 360 bps shortly following the Lehman Brothers bankruptcy on 15 September 2008 and remained elevated well into 2009 (Fig. 1).



LIBOR: Origins, Economics, Crisis, Scandal and Reform, Fig. 1 LIBOR began to display erratic behavior relative to other funding rates in the second half of 2008. Spreads to other funding rates widened drastically during the peak of the crisis, while LIBOR rates at times fell below what might be expected based on related rates

Rising spreads signalled the intensification of the crisis as liquidity and credit concerns drove interbank lenders to pare back funding while simultaneously demanding higher returns. Banks' inability to access funding in interbank markets fuelled perceptions of loss in creditworthiness, leading to a positive feedback loop that increased the credit risk component of LIBOR, ultimately driving spreads wider.

The reasons cited for elevated interbank rates stem from both the supply and demand sides. On the supply side, banks were unwilling to tie up funds for long periods of time due to balance sheet uncertainty brought about by the blossoming subprime ordeal (term liquidity risk). Conversely, this fear of funding instability drove the same banks to demand more long-term funding for liquidity purposes. Burgeoning demand chasing a shrinking supply of interbank funds, compounded by perceived increases in credit risk arising from subprime sectors, drove up LIBOR rates to new heights. Furthermore, the shifts in supply and demand noted above apply most conspicuously in longer-term transactions, meaning that as past funding matures, they are replaced with shorter-term contracts that are more susceptible to

rollover risk for the borrower. These movements in tandem negatively impact credit fundamentals for the financial institutions in question, which further drives up LIBOR rates through the credit risk component (Wrightson ICAP 2007).

Scandal

In April 2012, LIBOR came under heavy public scrutiny due to controversy over individual panel bank submissions during the height of the financial crisis. Allegations arose that banks had purposefully underreported their borrowing costs by significant amounts in order to project financial strength amidst market uncertainty. In addition, banks were alleged to have manipulated the rate to realise gains on LIBOR-based contracts. Whereas financial strength can be signalled by underreporting one's own submission, gains in LIBOR-based contracts require concerted action by multiple banks to influence the final fixing.

Though many banks were allegedly involved in misreporting, the most prominent to have reached settlements to date are Barclays, UBS, RBS, and Rabobank. CFTC probes ultimately concluded that the firms had acted in violation of the Commodity Exchange Act's false reporting provision (Gensler 2012). In addition to paying a settlement of $453.6 million to US and British financial authorities ($200 million to the CFTC, $160 million to the Department of Justice and $93.6 million to the UK Financial Services Authority), Barclays also lost a number of senior executives in the aftermath of the scandal, including CEO Robert Diamond, who resigned on 3 July 2012. UBS settled on 19 December 2012 for $1.52 billion ($700 million to the CFTC, $500 million to the Department of Justice, $259 million to the UK Financial Services Authority and $64 million to the Swiss Financial Market Supervisory Authority), RBS on 6 February 2013 for $612 million ($325 million to the CFTC, $150 million to the Department of Justice and $137 million to the UK Financial Services Authority) (WSJ 2013), and Rabobank on October 29, 2013 for $1.07 billion ($475 million to the CFTC, $325 million to the Department of Justice, $170 million to the U.K. Financial Conduct Authority, and $96 million to Dutch Authorities) (Bray 2013). Rabobank Chairman Piet Moerland also resigned as a result of the scandal.

Wrightson ICAP's weekly newsletter from 3 September 2007 may have been the first to publicly draw attention to the low level of LIBOR fixings. However, their analysis did not conclude that manipulation was the culprit, but instead settled on a dearth of interbank activity and the stickiness of official fixings to explain the observed divergence in rates (Wrightson ICAP 2007). The mainstream media did not catch on until a series of *Wall Street Journal* articles in 2008 exposed the possibility of targeted misquotes (Mollenkamp 2008; Mollenkamp and Whitehouse 2008b) (see Fig. 2 for a timeline of the LIBOR scandal). The journalists raised two possible motives for misreporting. The first involved a bank's desire to keep its submissions low in order to project an image of soundness. Robust capitalisation would help fend off media and market speculation surrounding funding difficulties during the height of the crisis. The second motive involved falsification with the expressed intent of benefiting the bank's derivatives positions. While early reports placed greater emphasis on the former argument instead of the latter, the authors provided no conclusive statistical evidence of actual manipulation. A subsequent *Financial Times* article by former Morgan Stanley trader Douglas Keenan suggested that LIBOR manipulation had been a fixture of financial markets as early as 1991 (Keenan 2012).

On the regulatory side, the Federal Reserve Bank of New York had first become aware of manipulative activities in 2007, with senior Federal Reserve officials being briefed by early 2008 (Reuters 2012). Correspondence between New York Fed President Tim Geithner and Bank of England authorities around the topic of LIBOR took the form of a 1 June 2008 email memo putting forth 'Recommendations for Enhancing the Credibility of LIBOR'. These recommendations included the establishment of best practices for calculating and reporting rates, the expansion of the USD LIBOR panel to a broader set of banks, the addition of a second USD LIBOR

LIBOR: Origins, Economics, Crisis, Scandal and Reform, Fig. 2 Timeline of the LIBOR scandal

fixing to reflect transactions that occur during US market hours, the specification of the transaction size at which submitted rates are applicable, the reduction of the number of maturities reported, and the elimination of incentives to misreport (FRBNY 2012).

During the course of investigation, Barclays pointed out that allegations of rate fixing during the peak of the crisis were inconsistent with the fact that its submissions were often in the top quartile of survey responses and thus dropped in the calculation of the interquartile mean. It is important to note, however, that misreporting did not imply that the individual LIBOR submissions were consistently lower than those of competitors, but rather that submissions were lower than the bank's true cost of funding in the interbank market. Barclays, as well as any financial institution, could misreport and still have rates among the highest submitted because of its borrower risk profile. The system's design, in which rate quotes are provided by market participants who hold large financial positions indexed to LIBOR, introduces an inherent conflict of interest (Ellis 2011). Net creditors benefit from higher fixings, while net debtors benefit from misquotes in the opposite direction. Alhough rate calculation via a trimmed mean reduces the market impact of each individual submission, collaboration among panel banks can still result in meaningful divergences from true rates.

While statistical evidence of wrongdoing by banks, both in isolation and in tandem, remains difficult to pinpoint even today, internal communications unearthed during the probes proved instrumental in showing purposeful intent to misreport. The CFTC uncovered documents showing that Barclays' traders requested specific actions from those in the bank responsible for LIBOR survey submissions. Manipulation ran rampant across multiple currencies and tenors for the expressed intent of benefiting the bank's

proprietary trading positions. The CFTC also uncovered a management directive to 'keep LIBOR submissions lower to protect Barclays' reputation' (Gensler 2012).

LIBOR's divergence from related funding rates – including effective federal funds, repos and Treasuries – raised warning flags for a market already unnerved by early subprime mortgage fears. Signs of rate tampering, however, were most clearly demonstrated in movements of Credit Default Swap (CDS) prices. The price of CDS reflects the cost of insuring against the default of the underlying institution, and heightened fears of insolvency reflected in rising prices should in theory be mirrored by increases in a firm's cost of funding in the interbank market. Rate submissions by the individual panel banks, however, failed to keep pace with CDS market activity, prompting questions from market observers.

Statistical evidence of reference rate manipulation has been limited. Abrantes-Metz et al. (2008) build on the methodology used in the original WSJ article to tease out suspicious patterns in the data, though they are also unable to definitively find evidence of manipulation. The markers they identify are of data patterns inconsistent with what is expected under normal market functioning, though manipulation does not necessarily entail the creation of these markers, nor do these markers necessarily imply the existence of manipulation. Brousseau et al. (2009) show that strong statistical relationships among various rates that existed prior to the Lehman collapse disappeared in the aftermath of the failure, though they stop short of attributing the disappearance to LIBOR manipulation rather than to the exogenous shock of the crisis itself. Ellis (2011) summarises the key empirical findings, highlighting in the process the dearth of concrete evidence for rate manipulation. Snider and Youle (2012) are perhaps the least reticent in their diction. They report that rationalising banks' LIBOR submissions proved difficult in light of data from other currencies and measures of funding cost. The positive spread between Eurodollar bid rates and LIBOR from August 2007 to mid-2011 generally ranged from 10–40 bps and is reflective of anomalous market conditions, as offer rates should generally exceed bid rates in markets of similar financial products (Fig. 1). Furthermore, they suggest significant financial incentives to underreport actual borrowing costs, citing their statistical analyses that confirm the existence of frequent manipulation. Kuo et al. (2012) are more tempered in their assessment, discussing many potential factors that may have caused the rate divergences of roughly 30 basis points during the crisis peak.

One feature of survey design that garnered heavy attention is the identity of the hypothetical interbank borrower. During the crisis, there existed protracted periods when a large gap existed between LIBOR and EURIBOR for the US dollar, even though both rates target the same funding conditions. While LIBOR asks each respondent the rate at which the bank itself can borrow, EURIBOR takes a more high-level approach by asking about the funding ability of the average panel bank. The benefit of the latter methodology is to better approximate the true rate of borrowing by dampening the psychological impact of overconfidence. This documented effect suggests that a majority of the banks surveyed would think that they are above the median in funding ability, and as a result drive the rate fixing below its true value in the aggregate. On the other hand, if the psychological impact of the differing survey designs were not material, then LIBOR's persistently low volatility relative to EURIBOR would cast further doubt on the rate's credibility (Gensler 2012).

The lack of conclusive results is further belied by criticism of the methods used to test for manipulation. Michaud and Upper (2008) suggest that analyses comparing LIBOR submissions to other publicly disclosed costs of funding are not able to disentangle liquidity premia from credit risk, making comparison among inherently different funding rates difficult to justify. They hold the opinion that liquidity, or the lack thereof, played a greater role in individual banks' borrowing rates than perceived credit quality. Gefang et al. (2010) similarly demonstrate that the widening of the LIBOR-OIS spread during the financial crisis was more reflective of illiquidity than credit concerns, but that the importance of the two

competing risks depended on the location within the term structure. The statistical methods used in distinguishing liquidity effects from counterparty credit risk have come under question. A BIS study took the more optimistic angle that the divergence in comparable market interest rates, while unusually large, was a product of design rather than evidence of tampering. Differential influences due to credit quality and liquidity likely drove the wedge between interbank rates without necessitating manipulation on the part of individual banks. Differing methods of dealing with outliers also contributed to the misalignments observed in market rates (Gyntelberg and Wooldridge 2008).

Although manipulation appears to have been a remnant of the past, investigation into wrongdoing is far from over. Media reports indicate that Deutsche Bank, Rabobank and ICAP are nearing settlements with US and UK regulators, while others, including Citigroup, likely remain on the ropes. More than 40 private lawsuits against the LIBOR panel banks have surfaced in the scandal's aftermath, with plaintiffs ranging from individual bondholders to cities like Baltimore and Philadelphia (McCoy 2013). These suits have met with limited success in the legal arena as large portions of their claims have been struck down (Raymond and Mollenkamp 2013). Estimates of total potential settlements to be paid by LIBOR panel banks range from $8 billion to $88 billion (Gongloff 2012).

What started out as the LIBOR scandal has not been confined to the one rate or the one market. Regulatory inquiries have abounded amidst heightened sensitivities in the post-crisis environment. EURIBOR has experienced similar rate manipulation allegations, while several banks are under investigation for manipulative practices in the energy, commodity and foreign exchange markets.

Repair and Reform, or Replace

Financial regulatory bodies across the world, including the International Organization of Securities Commissions (IOSCO) and BIS, have joined in a coordinated effort toward reference rates reform in the wake of the LIBOR scandal. At the heart of these deliberations sits the Financial Stability Board (FSB), an international body established in 2009 to oversee global financial system reform. The FSB has convened an Official Sector Steering Group composed of central bankers and other regulators to 'coordinate consistency of reviews of existing interest rate benchmarks'. It has similarly convened a Market Participants Group to represent private sector interests and address issues that may arise in implementation and transition (FSB 2013).

One potential upside of the LIBOR scandal is that it has provided the political impetus to re-examine the general structure of reference rates. A decline in unsecured term interbank activity following the financial crisis and a gradual shift toward reliance on secured funding raises the question of whether a LIBOR-like rate, even if equipped with ample governance, is appropriate going forward. The move toward central clearing of derivatives mandated by the Dodd–Frank Act further reduces the economic relevance of reference rates with significant counterparty credit risk built in. Limiting derivative exposures to a small number of central counterparties (CCPs) drastically reduces the interconnectivities among financial institutions, thereby shielding the system from contagion should isolated defaults occur (BIS 2013). CFTC Chairman Gary Gensler has pointed out that the interbank market itself has changed dramatically since the 1980s, when LIBOR was first popularised. Interbank unsecured funding has been gradually falling out of favour among market participants, particularly in the aftermath of the financial crisis, as capital and liquidity rules were put in place that effectively disincentivised this form of lending. Term funding has also shifted toward the shorter end of the spectrum, placing tension on market depth among the longer LIBOR maturities (Gensler 2012). It remains to be seen whether these changes in the interbank market are now permanent fixtures of global finance or temporary responses to the anomalous macroeconomic environment.

The numerous questions facing policymakers today surround key attributes of the desired

reference rate. Should it be structured like LIBOR to reflect bank credit risk, or should it be conceived as a risk-free rate in the vein of OIS? Should it remain an uncollateralised rate or reflect collateralised lending? Should it be constructed as a single rate or as a composition of multiple rates? Should it be quoted for a range of maturities or solely reported on an overnight basis? Should it be calculated using terms on actual market transactions or rely on discretionary submissions?

Regardless of the answers to the above, regulators are still tasked with managing the continuity risk surrounding existing financial contracts. Any substantive overhaul of reference rates could entail significant legal complications involving the reference rate cited in legacy obligations. Pricing discontinuities and operational difficulties within back offices could pose potentially high costs. Inefficiencies and costs stemming from potential private party lawsuits dealing with legacy LIBOR contracts are not insignificant concerns. One potential solution for legacy contracts is to continue management and reporting of the traditional LIBOR-based rates until all contracts have effectively matured or dissolved. One potential drawback of this approach is that market adoption of the new reference rate(s) might face stronger resistance with LIBOR still in existence.

On the other hand, once transition to the new regime has taken place, clear positive externalities are realised in the use of the same single reference rate. Network effects suggest that individual market participants benefit in a nonlinear fashion from the total number of users. Adoption of a single reference rate entails greater liquidity and maximises opportunities to trade and hedge against financial instruments tied to that rate; liquidity and market depth concerns would be all but eliminated. Such scale benefits would be harder to realise within a multi-reference rate regime, although risk diversification among numerous rates could prove beneficial should further shortcomings be discovered in any one of the rates. One further issue that comes into play is that of coordination. Heavy path dependency in the adoption process, akin to LIBOR's historical development, suggests a prominent role for policymakers. What is generally viewed as the socially optimal outcome may not be able to achieve critical mass if the adoption process for this public good is undertaken by the private sector in isolation.

One of the first official responses tackling the LIBOR issue came from the Financial Services Authority (FSA) in the form of the Wheatley Review. (The FSA was abolished effective April 1, 2013, with its duties split between the Prudential Regulation Authority and the Financial Conduct Authority. Martin Wheatley heads the latter agency.) The report highlighted the thinness of the market for a number of currency–maturity pairs, a trait that has persisted long past the crisis peak. It is striking to note that even the USD LIBOR, the most liquid of the 10 LIBOR currencies, suffers from this lack of market depth, as more than half of the 15 quoted maturities have reported little to no trading activity in recent years. The report proposes cutting out illiquid currency–maturity pairs and focusing instead on markets with sufficient trading data to support a transaction-based approach even in non-normal times. Moreover, the review concluded that transactions data should be explicitly used to corroborate discretionary submissions, without proposing that actual transactions be used in calculating the LIBOR fixing. It is further proposed that LIBOR oversight be transferred from the BBA to a government-sponsored administrator with statutory authority to bring about greater transparency and credibility. To combat the incentive to underreport funding costs and hence project an image of stability, the Wheatley Review recommends that bank-level submissions be published with a 3-month lag. Delayed public disclosure of component rates will also help repress rumours of changes in creditworthiness. The number of banks in the reporting panel should also be expanded to mitigate the effect of misreporting. Overall, public response to the Wheatley report has been positive. Rather than suggesting a complete overhaul of the system, the report seemed more focused on reforming the way in which the rate was administered (HM Treasury 2012; Wrightson ICAP 2012). The BIS report arrives at many similar

recommendations to the Wheatley Review, including increased usage of transactions data. Where the two reports differ is that the former pushes in particular for increased transparency in those markets where reference rates are derived, and encourages the development of alternative reference rates with minimal credit risk components (BIS 2013).

Other proposals for repairing rather than replacing LIBOR abound. One option that has gained traction is to convert LIBOR into a transaction-based rate whereby a weighted average of actual rates is used to calculate the fixing. Proponents of this approach view it as a quick, low-cost method to restore the integrity of the reference rate, while critics caution about the potential for heightened volatility. Lack of market liquidity for less widely used currency–-maturity pairs, especially during times of stress when interbank markets freeze, has been cited as an important stumbling block. Using Fedwires Funds Service inferred interbank transactions, Duffie et al. (2013) find that USD interbank volumes are concentrated at 1, 3 and 6 month maturities, and that there is a moderate flow of new transactions, even during the 2007 to 2009 crisis period. The authors also conclude that usage of sampling windows makes a transaction-based approach feasible even during times of market illiquidity. IOSCO guidance in this regard settles on the principle that a benchmark should be "anchored in an active market having observable, bona fide, arms-length transactions" (IOSCO 2013). The phrasing purposefully sidesteps the exclusive requirement for transactions data in determining benchmark values, allowing administrator discretion in using ancillary market data for supplementary purposes should the need arise. (The interplay between EONIA and EURIBOR mentioned earlier can prove instrumental for policymakers in discussing the merits of a transaction-based approach to LIBOR reform. Drastic volatility in the spread between the two rates can signal misreporting during adverse financial climates and thus encourage the adoption of a transaction-based measure.)

In 2008, Citigroup's Scott Peng suggested a new NYBOR rate that would complement the controversy-laden LIBOR going forward. This rate would be calculated in much the same way as LIBOR, but be based solely on NY banks' cost of funds (Mollenkamp and Whitehouse 2008a). The New York Funding Rate (NYFR) came into existence in June of 2008, with rates published daily by interbank broker ICAP (Wrightson ICAP 2008a, b; Kuo et al. 2012).

The NYFR survey was conducted at 9:30 a.m. NY time, with calculation and publication of the fixing around 10:00 a.m. Rather than an offered rate, NYFR would ask for the mid-rate and only for the 1- and 3-month maturities. One further improvement on its ideological predecessor is that NYFR, like EURIBOR, asks for the rate at which a representative bank would likely be able to borrow, rather than the rate at which each respondent is individually able to borrow. Furthermore, the individual rate submissions would be published each day without accompanying identifying information on the respondent. NYFR also reflects broader market conditions for wholesale unsecured funding rather than just interbank deposits, extending the pool of potential lenders and instruments. Finally, NYFR began with a daily required minimum of 24 panellists, with the top and bottom six dropped and the remaining 12 averaged to produce the fixing. Gradual declines in reporting by banks forced ICAP to reduce the threshold to 16, then 12, institutions. On 3 August 2012, ICAP ceased to publish NYFR altogether due to an inability to meet its own survey response standards.

Coulter and Shapiro (2013) also attempt to transform LIBOR by positing a new committed-quote framework to address current shortcomings. Firstly, bank submissions would be based on actual transactions if available. In the absence of borrowing data, suspect submissions can be called into question by other panel banks. Third parties can then confirm willingness to lend at the rate in question, or confirm the whistleblower's allegations of misreporting.

Those in favour of replacing LIBOR altogether have rallied behind Gary Gensler. The Overnight

Index Swap (OIS) rate has been put forth as a leading candidate. 2010 witnessed the adoption of OIS rates by the London Clearing House and ICAP to discount various derivatives contracts (Brousseau et al. 2012). Some large investment banks have also joined the movement to discount payments on financial contracts using expected compounded overnight rates to mitigate the reliance on reference rates with a significant credit risk component (BIS 2013; Tett 2008). However, longer term OIS rates including 1-month and 3-month are not yet mainstream among market participants.

General collateral (GC) repo rates have also been proposed as a possible complement to the credit-risk dominated unsecured LIBOR. This proposal would use the General Collateral Finance Repurchase Agreement Index (GCFs Repo Index) in place of LIBOR, with the intent that the transaction-based index would better reflect true objective funding costs, demonstrate stronger resilience to illiquidity under market stress, and more effectively fend off attempts at manipulation due to central clearing. The index is calculated as the weighted average interest rate paid on overnight GCFs repo transactions, which are by definition fully collateralised by US Treasury securities, non-MBS agencies and agency MBSs. A key advantage of this approach in implementation is that no new administrative agency would need to be established for oversight purposes, as the Depository Trust & Clearing Corporation (DTCC) currently calculates the index and could continue in this role with minimal interjection. Furthermore, repo contracts are known to be an important wholesale funding source for large banks, and the regulatory reforms that have already taken place to address shortcomings in triparty repo markets make usage of the GCFs Repo Index as a reference rate even more appealing. Although the DTCC only began publishing the index in November 2010, the product to date has shown none of the shortcomings that have crippled LIBOR (DTCC 2013).

At an even more basic level than the GCFs Repo Index, Treasury rates themselves have been put forth as a potential replacement for LIBOR for many of the same reasons. The market for US Treasuries is likely the most liquid in the world, even under financial duress. Moreover, Treasury constant maturity rates were heavily used as a reference rate for ARMs prior to the popularisation of LIBOR, and in fact are still referenced by many ARMs today (Schweitzer and Venkatu 2012). The possibility of replacement using a combination of several rates has also been discussed.

As of September 2013, many of the proposed changes for reforming LIBOR have already been put in place. Five less frequently traded currencies have been discontinued (NZD, DKK, SEK, AUD, CAD), while the five that remain now only report the 1-day, 1-week, and 1-, 2-, 3-, 6- and 12-month maturities. The total number of currency–maturity fixing pairs has been reduced from 150 to 35, with the possibility for further consolidation in the future. LIBOR submissions from individual banks now experience a 3-month delay in publication, effective as of 1 July 2013. Finally, keeping in line with the Wheatley Review proposal, the BBA was relieved of its duties in administering LIBOR. NYSE Euronext won the competitive bid for LIBOR for a nominal price of d1. The deal was announced on 9 July 2013, although the actual transfer of duties is expected to occur in early 2014.

See Also

- ▶ Banking Crises
- ▶ Banking Industry
- ▶ Credit Crunch Chronology: April 2007–September 2009
- ▶ Natural Rate and Market Rate of Interest
- ▶ Regulatory Responses to the Financial Crisis: An Interim Assessment
- ▶ Subprime Mortgage Crisis

Bibliography

Abrantes-Metz, R., M. Kraten, A. Metz, and G. Seow. 2008. LIBOR manipulation? *Journal of Banking and Finance* 36: 136–150.

Acharya, V., and D. Skeie. 2011. A model of liquidity hoarding and term premia in inter-bank markets. *Journal of Monetary Economics* 58(5): 436–447.

Bank for International Settlements. 2013. *Towards better reference rate practices: A central bank perspective*. Working Group report, Bank for International Settlements.
Bray, C. 2013. Dutch bank settles case over LIBOR deceptions. *New York Times*, Dealbook October 29, 2013.
Brousseau, V., A. Chailloux, and A. Durré. 2009. *Interbank offered rate: Effects of the financial crisis on the information content of the fixing*. Working Paper series 2009-ECO-10.
Brousseau, V., A. Chailloux, and A. Durré. 2012. Fixing the fixings: What road to a more representative money market benchmark? *IMF Working Paper* 13: 1.
Chen, J. 2013. LIBOR's poker: Interbank borrowing costs and strategic reporting. *UC Berkely Haas School of Business Working Paper*.
Coulter, B., and J. Shapiro. 2013. A mechanism for LIBOR. *University of Oxford Saïd School Business School Working Paper*.
DTCC. 2013. A primer on the DTCC GFC Repo Index. 14 September. Available at: http://www.dtcc.com/news/newsletters/dtcc/2012/sep/primer_gcf_repo_index.php
Duffie, D., D. Skeie and J. Vickery. 2013. A sampling-window approach to transactions-based LIBOR fixing. *Staff Report* No. 596, Federal Reserve Bank of New York.
Ellis, D.M. 2011. *LIBOR manipulation: A brief overview of the debate*.. FTI Consulting.
Federal Reserve Bank of New York. 2012. June 1, 2008: Timothy F. Geithner e-mail to Mervyn King, copying Paul Tucker, with attached 'Recommendations for Enhancing the Credibility of LIBOR', *New York Fed responds to congressional request for information on Barclays – LIBOR matter*. Available at: http://www.newyorkfed.org/newsevents/news/markets/2012/Barclays_LIBOR_Matter.html
Financial Stability Board (FSB). 2013. Meeting of the Financial Stability Board in Basel on 24 June.
Gefang, D., G. Koop, and S. Potter. 2010. Understanding liquidity and credit risks in the financial crisis. *Journal of Empirical Finance* 18: 903–914.
Gensler, G. 2012. *Remarks of chairman gary gensler, European Parliament, Economic and Monetary Affairs Committee*. Brussels, Belgium, September 24.
Gongloff, M. 2012. LIBOR scandal's potential costs exploding to $88 billion or more. *The Huffington Post*, August 27.
Gyntelberg, J., and P. Wooldridge. 2008. Interbank rate fixings during the recent turmoil. *BIS Quarterly Review*.
H. M. Treasury. 2012. *The wheatley review of LIBOR: Final report*. H. M. Treasury, London.
IOSCO. 2013. *Principles for financial benchmarks*. Madrid: IOSCO.
Keenan, D. 2012. My thwarted attempt to tell of LIBOR shenanigans. *Financial Times*, July 27.
Kuo, D., D. Skeie, and J. Vickery. 2012. A comparison of LIBOR to other measures of bank borrowing costs. *Work in progress*, Federal Reserve Bank of New York.
McAndrews, J., A. Sarkar, and Z. Wang. 2008. The effect of the term auction facility on the London Inter-Bank Offered Rate. *Staff Report* No. 335, Federal Reserve Bank of New York.
McCoy, K. 2013. 13 banks sued in rate-rigging case; federal credit union regulator says banks manipulated LIBOR. *USA Today*, September 25.
Michaud, F.-L., and C. Upper. 2008. What drives interbank rates? Evidence from the LIBOR panel. *BIS Quarterly Review* 3: 47–58.
Mollenkamp, C. 2008. LIBOR fog: Bankers cast doubt on key rate amid crisis. *Wall Street Journal*, April 16.
Mollenkamp, C., and M. Whitehouse. 2008a. LIBOR hits US borrowers. *Wall Street Journal*, April 23.
Mollenkamp, C., and M. Whitehouse. 2008b. Study casts doubt on key rate. *Wall Street Journal*, May 29.
Raymond, N., and C. Mollenkamp. 2013. Banks score major win in private LIBOR suits. *Wall Street Journal*, March 29.
Reuters. 2012. New York Federal Reserve knew about LIBOR rate-fixing issues as far back as 2007 and proposed changes but were ignored. *Daily Mail*, July 10.
Ridley, K., and H. Jones. 2012. A Greek banker spills on the early days of the LIBOR and his first deal with the Shah of Iran. *Reuters*, August 8.
Schwarz, K. 2010. Mind the gap: Disentangling credit and liquidity in risk spreads. *Working Paper*. University of Pennsylvania Wharton School of Business.
Schweitzer, M., and G. Venkatu. 2012. Alternatives to LIBOR in consumer mortgages. *Economic Commentary*, Federal Reserve Bank of Cleveland.
Smith, J. 2012. The term structure of money market spreads during the financial crisis. *NYU Stern School of Business Working Paper*.
Snider, C., and T. Youle. 2012. The fix is in: Detecting portfolio driven manipulation of the LIBOR. *University of Minnesota Working Paper*.
Taylor, J., and J. Williams. 2008a. A black swan in the money market. *NBER Working Paper* No. 13943.
Taylor, J., and J. Williams. 2008b. Further results on a black swan in the money market. *Working Paper*, Stanford University.
Tett, G. 2008. Lenders examine LIBOR alternatives. *Financial Times*, April 16.
Thornton, D. 2009. What the LIBOR-OIS spread says. *Economic Synopses*, No. 24, Federal Reserve Bank of St. Louis.
Wall Street Journal. 2013. *The LIBOR investigation*. Available at: http://stream.wsj.com/story/the-libor-investigation/SS-2-32262/. Accessed 21 Sept 2013.
Wrightson ICAP. 2007. LIBOR: *Twin Conundrums*. September 3.
Wrightson ICAP. 2008a. *Fed Policy*. May 2.
Wrightson ICAP. 2008b. *Federal reserve data*. June 11.
Wrightson ICAP. 2012. *The final report of the wheatley review of LIBOR*. October 1.

Libya, Economics of

Barry Turner

Keywords
Libyan dinar; Millemes; UN Sanctions

JEL Classifications
O53; R11

Overview

Libya enjoyed solid growth performance in the first decade of the century. In 2003 UN sanctions were lifted after Libya admitted complicity in the 1988 Lockerbie aircraft bombing. In 2004 the USA lifted almost all of its unilateral sanctions after Libya agreed to give up its nuclear weapons programmes.

In 2007, petroleum and natural gas contributed 71.6% to GDP; followed by public administration, defence and services, 6.9%; finance, insurance and real estate, 6.2%; and construction, 4.3%.

Libya suffered a recession in 2009 following the global economic downturn. As a result, there was a greater contribution from non-oil industries following increased government spending and liberalization of the trade, tourism and service sectors.

GDP contracted by 61% in the year after the February 2011 revolution. The interim government approved a budget worth almost US$50bn. for reconstruction and the election of the General National Congress in July 2012 prompted an estimated 76.3% boost in GDP that year. The hydrocarbon sector accounted for about 95% of total fiscal revenue in 2011–12 while oil and gas account for 2% of total employment. However, production encountered severe disruptions in 2013 as a result of militia blockades of oil facilities.

Private sector growth and economic diversification are essential for long-term stability. Corruption and unemployment must also be addressed, with the youth jobless rate estimated at 50%.

Currency

The unit of currency is the *Libyan dinar* (LYD) of 1000 *millemes*. The dinar was devalued 15% in November 1994, and alongside the official exchange rate a new rate was applied to private sector imports. Foreign exchange reserves were US$29,315 m. in June 2005. Total money supply in May 2005 was 11,552 m. dinars. There was inflation of 2.5% in 2010, increasing to 15.9% in 2011.

Budget

In 2008 revenues totalled 72,741 m. dinars and expenditures 44,115 m. dinars. Oil accounts for 88.6% of government revenues.

Performance

The economy contracted by 2.3% in 2009 but grew by 4.2% in 2010. Total GDP in 2009 was US$62.4 bn. The Libyan armed conflict led to the economy contracting by 59.7% in 2011—the highest percentage decrease of any country that year.

Banking and Finance

A National Bank of Libya was established in 1955; it was renamed the Central Bank of Libya in 1972. The current *Governor* is Saddek Omar Elkaber. All foreign banks were nationalized by Dec. 1970. In 1972 the government set up the Libyan Arab Foreign Bank. The Agricultural Bank was set up to give loans and subsidies to farmers and to assist them in marketing their crops. Following the popular uprising against Col. Gaddafi in 2011, international sanctions

were imposed on the central bank and its subsidiaries. In March 2011 rebels assumed control of the Central Bank of Libya, setting up a temporary headquarters in Benghazi. Following the overthrow of Gaddafi in August 2011, authority passed to the National Transitional Council.

A stock exchange was opened in Tripoli in March 2007.

See Also

- ▶ Energy Economics
- ▶ International Monetary Fund
- ▶ Islamic Economic Institutions
- ▶ Islamic Finance
- ▶ Oil and the Macroeconomy
- ▶ Organization of the Petroleum Exporting Countries (OPEC)

This is an edited and updated version of the economic profile of this country that appears on The Statesman's Yearbook Online: http://www.statesmansyearbook.com/

Licensing of Copyright Works

Richard Watt
University of Canterbury, Canterbury, New Zealand

Abstract

The article discusses the microeconomic theory behind copyright licensing. It analyses and mentions (at the least) the general economics of copyrights and how the income that is generated along the value chain can be transferred and shared using contracts, issues related to risk and risk bearing in copyright licensing, issues related to market power, issues related to hold-up and essential inputs, and issues related to collective licensing arrangements. The flavour of the article is theoretic, and it is written at a level of upper undergraduate economics courses. It equips the reader with a good knowledge base for understanding the principle issues that are at play as far as copyright licensing is concerned.

Keywords

Licensing; Copyright; Economics; Efficiency

JEL Codes

D; D23; D45; D71; D86

Introduction

The underlying tenet of almost all economic analysis of markets is the study of transactions, under which an item is passed from one economic actor to another who values it more highly, normally in exchange for a sum of money. Such transactions are regulated by contracts – either explicit or implicit – which dictate the terms and conditions under which the exchange takes place. For the case of copyright works – the fruits of the intellectual abilities of creators, which are protected by copyright law – most of the relevant transactions are best described as licensing, as opposed to direct sale and purchase, and from an economics perspective this in turn leads to many interesting aspects of such transactions (see Caves 2000 for an excellent treatment of the real-world of contracting in the space of copyright protected goods). It is the purpose of the present article to discuss and analyse some of these aspects from the perspective of microeconomic theory. There are several papers that deal with licensing contracts for copyrights, and that point out several particular features of these contracts (see the symposium in *Review of Economic Research on Copyright Issues* vol. 3(1) of 2006 – available online at http://www.serci.org/rerci.html, and in particular the short survey in Watt 2006, and more recently Watt 2013).

Copyright is a "property right", similar in spirit to any other property right. Its primary function is to protect the free exercise of rights from interference by others. For example, the right to possess, to use, to copy, to alter, to sell, to rent, to gift, to

bequeath, and to exclude others are all protected under copyright. With those rights properly protected, they are then able to be exercised in exchange for monetary recompense, and this is what (in principle) provides the incentives for creative work.[1] In short, so long as the law provides a complete and exhaustive definition of the bundle of rights that the property implies, then the bundle can become the subject matter of contracts, individually or in sub-bundles. Copyright itself does not provide an incentive to create, but it is what allows meaningful contracts to be written, and those contracts provide incentives to create, and to undertake the risks that are present in the value chain between creation and final consumption.[2]

Perhaps the main difference between copyright and other property rights is the fact that generally the actual intellectual creation itself is never sold. The contracts only allow for access of one type or another to the creation, and in that sense, contracts involving copyright protected creative goods are much closer in spirit to rental contracts than outright sales. It is this feature that leads to these sorts of economic transactions being referred to as "licensing" arrangements.

It is important to understand the differences between a standard "sale and purchase" agreement, and a "licensing" arrangement. Under a sale and purchase, *all* property rights in the transacted item pass from the vendor to the vendee, forever (or at least until the rights are sold on to a new owner). The fee paid by the vendee to the vendor in a standard sale is normally simply referred to as the "price". On the other hand, under a licensing arrangement, only some rights are transferred and only for a limited time, which is why a licensing deal is much more like a rental contract than an outright sale – the licensor (normally the copyright holder, or his/her agent) allows the licensee (normally another creator, a publisher, a distributor, a broadcaster, etc.) access to a limited set of rights concerning the creative work in question, for a limited period of time. During the time period stipulated by the licensing contract, the licensee can make use of, and profit from, those rights that are specifically covered by the agreement. The fee paid by the licensee to the licensor in exchange for access to the rights covered by the arrangement is often called a "royalty", in reference to the fact that the amount of the fee normally varies with the earnings made by the licensee during the contracted period. At the end of the agreement period, all access to the rights is surrendered, and the copyright holder once again becomes the sole proprietor of all such rights. The only caveat to that is if the licensing agreement runs up to the end of the period of legal protection of the copyright in question, in which case the licensing agreement gives way to the copyright falling into the public domain, whereupon all of the corresponding rights are forever more legally accessible by anyone at all without cost. The limited time duration of copyright is a response to a social welfare criteria. Copyright needs only to last as long as it takes for the expected present value of licensing income to fully compensate the creator for his/her efforts. Any longer and the creator is over-compensated, any shorter and the creator is under-compensated. As soon as copyright is long enough for the work to be created in the first place, then extending it any further causes a social cost in terms of access to the work that is not offset by any social benefits in terms of what the creator offers.

To see how licensing contracts might work, take the example of a novel that is produced as a book for sale to readers (one can easily make the terminological substitutions for the example to be for a musical composition, or any practically other copyright protected creation). The first thing is to clearly differentiate between the physical book and the novel that is written in the pages of the book. The first (the book) is an item of pure physical property, and as such is not covered by copyright law at all, and which is normally transferred under standard sale and purchase agreements. The second (the novel) is intellectual property, which is covered by copyright. The original author, upon completion of the work, becomes the copyright holder in the novel, but

[1]The exception are the author's moral rights, which cannot be traded or licensed.

[2]The sorts of creative works here are such things as musical compositions, lyrics, novels, poems, and art forms. These would be presented to consumers embedded in musical CDs and files, movies, books, and paintings.

the author is not normally well placed to shift the novel along the chain of activities and production that lead to a book being read by a final consumer. So the author (perhaps through an agent) contracts with a publisher to produce the physical book containing the novel. At this point we would see the first licensing contract taking place – the author allows the publisher the right to reproduce the novel, and to on-sell copies of the novel in book form to retailers, or perhaps directly to consumers. Notice that not all rights in the work are transferred to the publisher. For example, the publisher would not be allowed to alter the work in any way. The publisher may also be restricted as to which countries the book can be released in, and possibly at which price the book will be sold. The publisher would not normally be allowed to contract out other adaptations of the work, say to a movie producer. And of course the licensing contract will often stipulate an end-date, beyond which the publisher can no longer continue to produce copies of the work. The payment received by the author – the licensing fee – may take many forms, but often it will involve an up-front (non-refundable) advance on future royalties, and a payment for each unit of the published book that is ultimately sold to a final consumer.[3] Interestingly, this sort of contract mimics exactly a deductible insurance contract, whereby the publisher is insuring the author (see Watt 2013).

Further contracts then occur along the value chain. The publisher will typically conclude a contract with retailers giving the license to distribute the book, which will normally stipulate a monetary payment back to the publisher for each book that is sold to a consumer (perhaps an individual, perhaps a library). Those contracts will also be of a licensing form, as the retailer will have a restricted set of permissible actions (perhaps even not having a free choice of retail price). And lastly, of course there is the final contract between the retailer and the final consumer, which is a sale and purchase of the book itself for a stipulated price, but which we can also understand as a licensing arrangement for the novel (the intellectual property), which allows the consumer the right to read, but not copy in any way, the novel that is printed in his/her book.

As can be seen by the above example, the entry point for actual money is when the final consumer purchases a copy of the book. All of the other contracts along the value chain simply serve to *share* the retail price received from the consumer among all of those parties that made the book possible – the retailer keeps a portion and passes the rest back to the publisher under the distribution license contract. The publisher keeps a portion and passes the rest back to the copyright holder (the author in our example) under the initial royalty contract. And if the author acts through an agent, then there is a further step with contracted payments between the publisher and the author. As such, what a licensing arrangement does is to share revenue among several players who jointly collaborate in some way to create a consumable item that is valued by final users. Depending on how the contracts are structured, the payments that flow from one participant to the next serve two main purposes; (i) to compensate participants for the time and effort dedicated to their part in the production process, and (ii) to compensate participants for sharing in the risks associated with the enterprise.

So long as the final demand for the end-product is uncertain or risky while the product is moving through the production chain (which is, essentially, always), risk is an ever-present element that the contracts need to take into account. The contracts need to be concluded *before* the amount of money that will be available to be shared is known. The structure of each licensing arrangement will reflect not only how valuable each participant is in the joint effort of producing the consumable product, but also how each of them will share in the risks of the venture. It is the need to share risk that leads to copyright licensing contracts often being characterised by "royalty"

[3]In principle, the copyright holder and the distributor will freely bargain over the payment terms in the contract. However, in practice publishers typically hold a much stronger bargaining position than do most authors, and so authors are only offered a take-it-or-leave-it standard contract. Some authors – those with established reputations – offer a much lower risk to publishers, and will be able to negotiate more favourable contracts for themselves. The issue of bargaining will be discussed in more detail below.

payments, that is, payments that depend in some way upon the final outcome achieved in the market where the end-product is sold to consumers.

It is probably a fair comment to say that economists have been much more interested in studying how licensing contracts share risk than in how they remunerate the value of productive efforts. Perhaps this is because "productive effort" may be understood to refer to actions that make the end-product more valuable, and to actions that improve the risk profile of the venture, and in many instances these two things are confounded. That is, a venture with a better risk profile is a more valuable one. For example, assume that an author of a successful novel has agreed, under a licensing contract, for a movie studio to make a film of the novel. At that point there is little that the author can do to affect the value and/or risks of the final film.[4] But the choices and investments that the studio makes around which actors it casts in the main roles and around how it goes about marketing the film can both be thought of as increasing the value of the end-product, and as reducing the risk that the film fails at the box-office.

The principle difficulty with analysing the risk sharing nature of licensing contracts is that, in theory, the theoretically optimal contract will share risk according to the aversion that each participant has to risk (their risk aversion), and this is in any real-world scenario private information to each party to the contract. All we can say in general is that the more risk averse is a given party, the less risk that party should carry in the final risk sharing arrangement. However, the optimal contract should also give participants the correct incentives to carry out actions that would have the effect of improving the risk profile of the venture, so long as they can do so at a low enough cost. In short, a licensing contract is a special type of principal-agent relationship (see Pérez Castrillo and Macho Stadler 2014).

[4]It is more normal for movie script writers to be paid by a flat fee, and not to participate in the future revenues of a film as might directors and principal actors.

The General Structure of Licensing Contracts and Risk-Sharing

Licensing contracts normally have at most two features, a fixed or constant component and a payment that is functionally related to the sales of the end-product, or a "royalty". Some licensing contracts may have only one of these two components (i.e. some may have only a fixed fee, others may have only a royalty and no fixed fee), others may have both components. Other licensing arrangements may not be related at all to the sales volume but rather to the profits or revenue of the intermediary (see, for example, Watt 2011). With that in mind, the licencing income that the copyright holder receives is given by a function of the type

$$L(X) = F + R(x)$$

where F is a fixed fee, $R(x)$ is a royalty payment that depends on the number of units sold, x.

Exactly how a licensing contract will balance the use of the fixed payment and the royalty option will depend on many things. We shall discuss some of the more interesting of them below. However, the most interesting element of such a "two-part tariff" licensing payment is the setting of an optimal royalty from the perspective of the copyright holder (see, for example, Besen and Kirby 1989, Varian 2000, and Watt 2000). If the royalty in a licensing agreement between a copyright holder and a publisher/distributor of the copyright work is an increasing function of sales, as would be normally expected, then it implies an artificial increase in the marginal costs of the distributor. This has two immediate perverse effects for the copyright holder's licensing income. First, if the increased marginal cost is passed on to the next agent in the distribution chain as an increased per-unit price, then we should expect lower sales to result. And under lower sales, the amount of royalty income due to the copyright holder can reduce rather than increase. Thus, it is a delicate question as to what the optimal copyright royalty should be, if the copyright holder sets the royalty with the objective of maximising royalty income. Second, if there is the possibility that the legitimately

produced copies of the copyright work might compete with pirate copies in the market, then the increased marginal cost of the producer of legitimate copies via a royalty element in the licensing contract increases the marginal cost disadvantage of the legitimate distributor relative to the pirate. This can clearly exacerbate piracy, which may come at the expense of legitimate sales, thereby also reducing the total licensing income of the copyright holder. Woodfield (2006) posits that it is even possible in such a setting that the optimal royalty payment for the copyright holder could become negative. In what follows, we will ignore these issues in order to focus our attention on other important features of royalty payments within licensing contracts.

It is important to notice is that the inclusion of a royalty only makes sense (in a single period model) if there is risk or uncertainty over the final amount of sales that will be realised. If x were not risky, that is, if it were perfectly known at the time the contract is written, then $R(x)$ would just be a constant number. That is, the contract $F + R(x)$ only contains fixed numbers, so the entire contract is effectively a fixed fee and it may as well only be expressed as a single fixed payment with no reference at all to x.[5]

If, on the other hand, there were a dynamic aspect to the situation, then a royalty can be included even if there is no risk. This was shown by Towse (2001), who posited that there is an interdependency relationship between the value of future works by a given author and the value of their existing works, through a reputation effect. Current success drives up the price of future works, and it also drives up the price of previously created works. If that happens, then it is efficient for the author to retain some financial interest in the sales of the existing works, and this can only be achieved with a royalty. Clearly, then, in such a dynamic setting, royalty payments as a part of licensing contracts serve to give authors an incentive to continue to work hard to output high quality works, which is undoubtedly in the public interest.

Going back now to the single period setting, if there is risk in regards the value of x, but if both participants are risk-neutral, then again a royalty is not needed. This is because if the participants are risk-neutral, they only care about the expected value of their payoff, and not its variance. So long as the expected value of x, denoted by Ex, is known, then the contract will read as $L = F + R(Ex)$, which again is a constant. So there is no need to reference the copyright holder's payment to the variable x at all, i.e. a fixed payment will always suffice.

So the only reason to include a variable royalty element, $R(x)$, is that x is a random variable when the contract is signed, and that the participants are risk-averse. In those cases (which are most, if not all, relevant real-world cases), the function that is chosen for $R(x)$ is important, as it acts as a risk sharing mechanism, and as an incentive mechanism. Above all, notice that if the royalty function is linear, that is if $R(x) = rx$ for a given number r, then the royalty payment rx is larger the larger is the outcome of x. In general, it is true that for risk to be shared efficiently, the royalty function should *not* be linear. That is, in general the value of the marginal license fee as x changes is not constant (see Alonso and Watt 2003). This can be captured by simply setting $R(x) = r(x)x$, that is, the per-unit payment $r(x)$ itself depends on x. Linear risk sharing, which is when r is independent of x, as it happens, is efficient only for cases in which each participant in the contract has "equi-cautious" utility functions of the hyperbolic absolute risk aversion class – these are utility functions that display linear risk tolerance with the same slope parameter (see, for example, Wilson 1968, Amershi and Stoeckenius 1983, Pratt 2000 and Gollier 2001). Included in this class of functions are the commonly assumed cases of constant relative risk averse utility (power utility and logarithmic utility), and constant absolute risk averse utility (negative exponential utility).

We can easily see why in general a fully efficient licensing structure will require that r be

[5]Often, a "fixed fee" payment is actually an upfront payment against future royalties, that is, the licensee pays a royalty advance. As noted earlier, this mimics exactly a deductible insurance contract under which the licensee insures the licensor. It thus reduces risk for copyright holders, and increases risk for copyright users (see Watt 2013).

different for different outcomes of x, by considering a very simple example in which we violate the "equi-cautious" requirement for linear risk sharing to be efficient. Assume that there are only two possible outcomes for x, say $x_1 > x_2$, and assume that the publisher is risk-neutral and that the creator is strictly risk-averse. From the general theory of risk-bearing, in such a case it is efficient for the risk-neutral party (the publisher) to bear all of the risk in the venture, or in other words, the creator should bear no risk at all. This will require that the creator gets the same income regardless of the outcome of x, so in this case the licensing contract must stipulate that $L(x_1) = L(x_2)$. But if the same royalty parameter (assuming it is strictly positive) is used when the outcome is x_1 as for when it is x_2, this becomes impossible since $L(x_1) = F + rx_1$ is always greater than $L(x_2) = F + rx_2$ when x_1 is greater than x_2. Clearly this case can be easily resolved by requiring that the royalty parameter is 0, so that the licensing contract only stipulates a fixed payment F. However, notice that there is actually another way to make the licensing income constant for the creator even with a "royalty" element in the contract, but it will involve the royalty payment parameter being different when the outcome is x_1 as when it is x_2. Specifically, if a positive royalty payment is desired for some reason, then for the case at hand (risk-neutral publisher and risk-averse creator) the royalty payment for outcome x_1 must be equal to that of outcome x_2, i.e. $r(x_1)x_1 = r(x_2)x_2$. Since we are assuming $x_1 > x_2$, it clearly becomes necessary that $r(x_1) < r(x_2)$, that is, the per-unit royalty is lower the higher is the level of sales.

The case with a risk-neutral participant is easiest to resolve by simply using a royalty parameter of 0, and having the entire licensing fee be given by an appropriate value of F. But what if both participants are risk-averse, but one is more risk-averse than the other? This can be thought of as a step away from the previous case, where the step is to make our previously risk-neutral participant now a little bit risk-averse. It can be shown that with two strictly risk-averse participants (even if one of them is nearly risk-neutral), both will participate actively in sharing the risk, but in such a way that the more risk-averse participant will undertake relatively less risk (Wilson 1968). So, since the creator now does not receive a constant licensing fee, and must share in the risk, it will happen that the licensing fee should be higher the higher is the outcome of x. In the case of x only taking on two values as before with $x_1 > x_2$, it now happens that the efficient licensing contract must satisfy $L(x_1) > L(x_2)$, that is, the license fee is increasing in x. Clearly now it must hold that $r(x_1)x_1 > r(x_2)x_2$. But notice that this says nothing about how $r(x_1)$ compares to $r(x_2)$. The former could be greater than, less than, or equal to the latter. All that can be said for sure is that the royalty payment must exist (or else the license fee would be a constant, and would not share risk), and it must be greater the greater is the outcome of x. In order to say any more, we would need to know the utility functions of the two participants.

In passing, notice that when both participants are strictly risk-averse, so that the license fee implies a greater payment to the creator the greater is the outcome of x, then the contract stipulates that the creator is rewarded more for a better sales outcome. This not only shares risk efficiently, but of course such a feature may also lead to the creator putting in more effort to creating a more valuable work in the first place. Here, we can understand that a more valuable work is perhaps one with a greater chance of being successful (i.e. achieving the high sales level x_1) and a lower chance of being a flop (and achieving sales of x_2). If the creator really does have the opportunity to, at a reasonable cost, put in more effort to create a more valuable work, then it would be efficient that the contract offered by the publisher also pays more for a better market outcome. Such a situation is known as "moral hazard" (see Pérez Castrillo and Macho Stadler 2014), and if we were only to observe that a licensing contract pays more to a creator the better are final sales, we would not (in principle) be able to know if that feature is present to give the creator incentives to create a more valuable work, or if it is there for risk sharing purposes, or both.

Interestingly, it is obvious that the greater is the payment to the creator, the lower are the retained earnings of the publisher. Imagine then that the publisher could also take more or less efforts to

facilitate a good sales outcome. Then there would be double moral hazard, and the contract would need to provide the publisher with the correct incentive to invest in the optimal level of effort to enhance the likelihood that the sales outcome is good. This would then imply a contract that pays the publisher relatively more for a good outcome and relatively less for a bad outcome, which in turn must imply exactly the opposite for the creator. How exactly such a double moral hazard scenario would end up being reflected in the final licensing contract terms will depend on (at least) the levels of risk aversion of the two participants, the degree to which their efforts contribute to the probability of a good outcome, and the costs that each must suffer to deliver efforts for a more valuable work. In the end, quite a complex task indeed.

Bargaining and the Relationship Between Copyright Law and Royalty Contracts

In the end, contracts are voluntarily entered into by the parties involved. Thus, we should expect that two things happen in regards contracts; (1) they should increase the utility of both parties to the contract above their utility without a contract (or at least, the utility of neither party should be reduced by the contract), and (2) the terms and conditions of the contract will be set as a result of a bargaining process (see Muthoo 2006 for a thorough treatment of how copyright licensing is shaped by bargaining processes). The first point seems obvious – since contracting is voluntary, no rational individual would ever conclude a contract that makes him/her worse off. Indeed, it is during the bargaining process alluded to as point (2) that the two parties will come to terms that ensure that neither of them are worse off by signing the contract. Given, then, that a contract is welfare enhancing, we can assume that by entering into such an agreement, some sort of surplus (albeit a risky surplus) is created that would not be created otherwise. If the contract is between a publisher and an author, then the contract provides a mechanism under which the author's novel is transferred into a consumable (and sellable) state (an actual book), and then marketed to potential purchasers. Any final book sale that happens is an addition to the surplus that is created by the contract. In a nut-shell, what the contract does is to dictate how any surplus that ends up being created is shared between the two contracting parties, and that sharing rule is what gives each the incentive to carry out whatever tasks are required of them under the contract as their part in surplus creation.

The standard microeconomic model of bargaining, as set out by Nash (1950), is entirely applicable to the general case of copyright licensing contracts. Nash's model provides a solution to any bargaining situation (between two negotiating parties), that is given by a share of the surplus that cooperation creates. As one might easily expect, the exact solution that the Nash model arrives at depends upon all of the parameters that are assumed at the outset to reliably describe the negotiating situation at hand. For example, the solution will depend upon (among any number of other things) what are the utility functions of each party (which in turn captures their risk aversions), what are the different contingencies for the value of created surplus and how (and at what cost) each player can affect these contingent values, what are the values of their outside options, what are the informational differences over the two players, and how patient each one is in the process (again, the reader is referred to Muthoo 2006 for a detailed discussion).

Here then, we can see that different sets of the basic underlying parameters that define the situation will result in different outcomes of the bargaining process, that is, in different ways in which surplus is shared, and perhaps even in different amounts of surplus to be shared. For example, as we have already discussed above, so long as the amount of shareable surplus is risky, then different risk bearing attitudes of the two contracting partners (i.e. different levels of risk aversion) will typically lead to different negotiated sharing rules.

Another principle component in the bargaining process for contracts concerning copyright goods are any legal constraints and restrictions on sharing that might be implied by copyright law. For example, in most jurisdictions the moral rights in a

copyright work cannot be contracted, so to the extent that the final surplus depends on the moral rights in the work, the moral rights holder enters the bargaining process in a different position (perhaps with different outside options) than does the other party. There are also many important instances of copyright tribunals stepping in and regulating the tariff that can be set for certain contracts for access to copyright material. This happens, for example, in the case of many musical broadcasting platforms (terrestrial radio broadcasts, online streaming, etc.) around the world. There are several theoretical justifications for such a restriction on free bargaining. First, in musical broadcasting, the music itself is an essential input, and so there is a fear that the bargaining process might take far too long before an agreement is reached, since the copyright holders are able to "hold-up" the process, effectively holding the distribution platform to ransom, until an agreement is finally reached in which the platform surrenders to a deal in which most of the surplus goes to the party controlling the essential input. The main problem with this, of course, is that in the meantime no music broadcasting takes place, so consumers are the losers in the whole process. Setting a tariff by regulation allows the contracts to be concluded much more quickly, and the music broadcasting service is then able to function earlier.

Second, there may also be a fear that the bargaining process might for some reason lead to a perceived unfair allocation of surplus, for example, out of a perceived imbalance between the contracting parties in terms of market power. It is easy to imagine that when a lone copyright holder negotiates with a large international publishing company, there could be an issue of grossly unequal bargaining powers, which can easily lead to contract terms that, while they must still offer the creator an incentive to participate in the current contract, may not provide him/her the appropriate incentives to embark upon new creative projects in the future. If that is the case, and under the (very reasonable) assumption that it is socially valuable for the dynamics of creation not to be interrupted, there could well be justification for intervention by authorities. This, exactly, is the economic rationale for the existence of copyright law in the very first place.

In short, legal parameters and definitions from copyright law end up shaping the contracts that are negotiated for licensing of copyright works, by affecting the base rules, the outside options, and the very bargaining powers of the parties to the contractual negotiation process. It is very interesting to try to consider exactly how this process works, and to what extent copyright law clauses that are designed to aid copyright holders in one instance may end up hindering other copyright holders in other instances. However theoretical and empirical work along this dimension is scant. At least in the empirical domain, changes in copyright law have been too few and far between to allow reliable statistical analyses of the effects upon licensing contracts, and upon measures of the output of copyright works.

Collective Licensing of Copyright

One of the most important features of much of licensing of copyright works is the fact that many works are licensed by copyright collectives rather than by individual copyright holders. A copyright collective is a group of copyright holders who license works of a similar nature (e.g. musical works), to the same set of users (e.g. music broadcasters and streaming services). The collective works by offering a licensing contract to all of the works together (the repertory), under what is known as a "blanket license" – a license to access any and all of the works (only those rights covered by the license – e.g. the reproduction right, or the mechanical broadcast right), as often as the user wants, for the duration of the license period, in exchange for a single fee. Economic theory suggests two important reasons why copyright collectives might be efficient, in spite of the clear monopoly power that they might have – transaction costs savings, and better risk management (see Watt 2015).

The standard economic theory of copyright collectives (see, for example, Hollander 1984, and Besen et al. 1992), posits that the foundational

aspect upon which a copyright collective forms is the existence of transaction costs (e.g. initial search costs for users and works, bargaining costs to settle on an agreeable licensing arrangement, costs of monitoring use and collecting the relevant royalties, and the costs of ensuring that the contract is respected). Since many users want to contract with a similar set of many copyright holders (and vice-versa), if the contracts are carried out individually the aggregate transaction costs multiply unnecessarily, with many contractual actions that generate costs simply replicating actions already carried out for a different contract. The transaction costs can be efficiently shared when copyrights are exploited together rather than separately. If the copyright holders join together into a unified group, and if all that is offered to users is a blanket license for access to the copyrights of all of the works together, then the transactions costs are greatly reduced, and the implied savings can be shared on both sides of the ensuing contract. This is similar in nature to why bus tickets allow riders to exit at a variety of stops, why gymnasium subscriptions allow users unlimited access to the facility, and why Microsoft Office contains programs that many of us will never end up using. All of those are examples of blanket licenses. When transaction costs are factored into the business model the costs of running a collective are sub-additive (average cost diminishes with the size of the collective). In such an environment, the theory of natural monopoly ensures that it is efficient that licenses are granted collectively rather than individually. Thus, the transaction costs theory argues for aggregation of copyrights into a single repertory to be licenced.

However, there is a second rationale for the existence of copyright collectives and blanket licensing, namely risk management. A copyright collective can be thought of as a type of firm that takes as its inputs the copyrights to individual works, that bundles these copyrights together, and that then licenses access to the bundle to users. The supply-side efficiencies then are related to aggregation efficiencies of scale (through the Law of Large Numbers), rather than to actual production of a new good. A second relevant feature is that the input suppliers (the individual copyright holders) are not paid a set price for surrendering their work to the repertory, but rather their payoff is a share of the proceeds from the sale of blanket licenses to users. In that way, the individual copyright holders are much more like shareholders, or owners, in the firm, and the collective is much closer in spirit to a "mutual" firm, or what economists often call a "syndicate".

At the forefront of the economics of syndicates (see Wilson 1968), is risk sharing. An "optimal risk sharing problem" in economics is normally formulated as follows:

> Given an uncertain environment with n possible states of nature, an aggregate payoff $X = (X_1, X_2, \ldots, X_n)$, and m agents, the problem is to divide X into (uncertain) shares x_j j=1,. . .,m, where $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$, such that for each state i, $\sum_j x_{ji} = X_i$, with x_j being acceptable for agent j, $j = 1,\ldots, m$.

For the case of a copyright collective, the aggregate payoff, X, is the total amount of royalty income from licensing the repertory to users, and in order to participate in a copyright collective risk sharing problem, each member (individual copyright holder) should have contributed to the collective's repertory the rights to their own (risky) copyrighted composition.

It is quite clear from the definition given above that a copyright collective is indeed a syndicate, and therefore the economic theory of syndicates is entirely applicable. A general outline of the problem of efficient risk sharing within a syndicate is presented in Gollier (2001), chapter 21, where the Pareto optimal risk sharing arrangements are also discussed. While undoubtedly a very important feature of copyright collectives, the issue of exactly how licensing income is shared among the members is somewhat parallel to, rather than at the heart of, the topic of licensing itself. Therefore this will not be covered in the present article, but rather the interested reader is referred to the analysis and discussion in Watt (2015). Suffice it to say that aggregation of many risky copyrights into a single bundle for licensing to users is very efficient from a risk-management perspective, and it allows the corresponding efficiency gains to be shared on both sides of the licensing contract (i.e. it will certainly have an effect on the licensing

terms as compared to what would occur without collective management of the repertory items).

In a somewhat related literature (see, for example, Bakos and Brynjolfsson 1999 and Bakos et al. 1999), it has also been shown that, in the relationship between suppliers and demanders of information goods, bundling is both economically efficient and profit enhancing. Since copyrights are pure information goods, then bundling into a single repertory is favourable for copyright holders. This efficiency occurs because bundling reduces the heterogeneity in user willingness to pay for individual titles (a bundle allows a sale at the average, rather than the minimum, willingness to pay). The theory of bundling is relevant to the licensing arrangements that will ensue between a copyright collective and the users of the repertory. The suggestion is that bundling will swing the bargaining power in favour of the collective, resulting in more favourable licensing terms than what would be achieved by copyright holders acting alone.

Conclusions

This article has considered and reviewed the microeconomic theory of licensing of copyright works. In general, there are several relevant areas of microeconomic theory that can be directly applied to licensing of copyright works – the theory of asymmetric information, the theory of optimal risk-sharing, bargaining theory, and the theory of aggregation – to name a few. Due to limited space, a few topics were either covered rather quickly or not covered at all (regulation of licensing arrangements, and licensing over international boarders, for example). Above all, the main points to understand are that licensing contracts for access to copyright works provide an incentive environment for creators to create, publishers to publish, and users to use. That is, licensing at the various points along the production-distribution-consumption chain is what allows a market to function, with the corresponding creation of social welfare. Licensing essentially dictates how both the final value that consumers place upon copyright protected goods, and the risks that are present in that final value, are shared among those individuals and firms that jointly produce the consumable item embodying the copyright protected work. Licensing contracts are also the mechanism that, together with an appropriately enforced copyright law, provides creative individuals with an incentive to continue to create content in the future, thereby ensuring that the whole process continues over time.

See Also

- ▶ Aggregation (Theory)
- ▶ Bargaining
- ▶ Risk Sharing

Bibliography

Alonso, J., and R. Watt. 2003. Efficient distribution of copyright income. In *The economics of copyright: Developments in research and analysis*, ed. W. Gordon and R. Watt. Cheltenham: Edward Elgar.

Amershi, A.H., and J.H.W. Stoeckenius. 1983. The theory of syndicates and linear sharing rules. *Econometrica* 51 (5): 1407–1416.

Bakos, Y., and E. Brynjolfsson. 1999. Bundling information goods: Pricing, profits and efficiency. *Management Science* 45 (12): 1613–1630.

Bakos, Y., E. Brynjolfsson, and D. Lichtman. 1999. Shared information goods. *Journal of Law and Economics* 42 (1): 117–156.

Besen, S., and S. Kirby. 1989. Private copying, appropriability and optimal copying royalties. *Journal of Law and Economics* 32 (2): 255–280.

Besen, S., S. Kirby, and S. Salop. 1992. An economic theory of copyright collectives. *Virginia Law Review* 78: 383–411.

Caves, R. 2000. *Creative industries: Contracts between art and commerce*. Cambridge, MA: Harvard University Press.

Gollier, C. 2001. *The economics of risk and time*. Cambridge, MA: MIT Press.

Hollander, A. 1984. Market structure and performance in intellectual property: The case of copyright collectives. *International Journal of Industrial Organization* 2 (3): 199–216.

Muthoo, A. 2006. Bargaining theory and royalty contract negotiations. *Review of Economic Research on Copyright Issues* 3 (1): 19–27.

Nash, J. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Perez-Castrillo, J., and I. Macho-Stadler. 2014. Copyright licensing under asymmetric information. *Review of Economic Research on Copyright Issues* 11 (2): 1–26.
Pratt, J. 2000. Efficient risk sharing: The last frontier. *Management Science* 46 (12): 1545–1553.
Towse, R. 2001. *Creativity, incentive and reward: An economic analysis of copyright and culture in the information age*. Cheltenham: Edward Elgar.
Varian, H. 2000. Buying, sharing and renting information goods. *Journal of Industrial Economics* 48 (4): 473–488.
Watt, R. 2000. *Copyright and economic theory: Friends or foes?* Cheltenham: Edward Elgar.
Watt, R. 2006. Licensing and royalty contracts for copyright. *Review of Economic Research on Copyright Issues* 3 (1): 1–17.
Watt, R. 2011. Revenue sharing as compensation for copyright holders. *Review of Economic Research on Copyright Issues* 8 (1): 51–97.
Watt, R. 2013. Copyright law and royalty contracts for copyright. In *Handbook of the digital creative economy*, ed. R. Towse and C. Handke. Cheltenham: Edward Elgar.
Watt, R. 2015. The efficiencies of aggregation: An economic theory perspective on collective management of copyright. *Review of Economic Research on Copyright Issues* 12 (1/2): 26–45.
Wilson, R. 1968. The theory of syndicates. *Econometrica* 36: 119–132.
Woodfield, A. 2006. Piracy accommodation and the optimal timing of royalty payments. *Review of Economic Research on Copyright Issues* 3 (1): 43–60.

Lieben, Richard (1842–1919)

Jürg Niehans

Keywords

Auspitz, R.; Consumer's rent; Gold standard; Lieben, R.; Mathematical economics; Reciprocal demand curves; Walras, L.

JEL Classifications

B31

Lieben was born on 6 October 1842, in Vienna; he died there on 11 November 1919. After studying mathematics and engineering sciences, he became a partner in the Jewish family bank and a respected member of the Viennese business community. In 1892 he advocated the adoption of a gold standard. He married late and had no children. He seems to have been of scholarly and artistic tastes, more contemplative than active.

Together with his cousin and brother-in-law Rudolf Auspitz, Lieben wrote the 'Researches on the Theory of Price' (1889), the only Austrian contribution to mathematical economics and one of the outstanding contributions in the last two decades of the 19th century. (This book is discussed in the dictionary entry on Auspitz, Rudolf.)

As a correspondent to the *Economic Journal*, Lieben provided a lucid summary of their views on consumer's rent (1894). After his collaborator's death he concluded the controversy with Walras by a complex three-dimensional analysis of reciprocal demand curves (1908), gracefully acknowledging their original misunderstanding.

Appropriate corrections were made in the French translation of the 'Researches' (1914). While it is impossible to separate Auspitz's and Lieben's contributions, these papers suggest that Lieben was more than a junior partner.

Selected Works

1887. (With R. Auspitz). *Zur Theorie des Preises*. Leipzig: Duncker & Humblot.
1889. (With R. Auspitz). *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot. French translation by Louis Suret, Paris: M. Giard & E. Brière, 1914.
1890. (With R. Auspitz). Reply. (To article by L. Walras in same publication). *Revue d'économie politique* 4.
1894. On consumer's rent. *Economic Journal* 4: 716–719.
1897. Ueber die weitere Ausdehnung des Wasserstrassennetzes in Oesterreich in der Zukunft. *Verbands-Schriften des Deutsch-Oesterreichisch-Ungarischen Verbandes für Binnenschiffahrt*.
1898. Indian currency. *Economic Journal* 8: 151–152.

1908. Die mehrfachen Schnittpunkte zwischen der Angebots- und der Nachfragekurve. *Zeitschrift für Volkswirtschaft, Sozialpolitik und Verwaltung* 17: 607–616.

Bibliography

Weinberger, O. 1931. Rudolf Auspitz und Richard Lieben. *Zeitschrift für die gesamte Staatswissenschaft* 91: 457–492.

Winter, J. 1927. *Fünfzig Jahre eines Wiener Hauses*. Vienna: F. Jasper.

Life Cycle Hypothesis

Malcolm R. Fisher

The life cycle hypothesis presents a well-defined linkage between the consumption plans of an individual and his income and expectations as to income as he passes from childhood, through the work participating years, into retirement and eventual decease. Early attempts to establish such a linkage were made by Irving Fisher (1930) and again by Harrod (1948) with his notion of hump saving, but a sharply defined hypothesis which carried the argument forward both theoretically and empirically with its range of well-specified tests for cross-section and time series evidence was first advanced in 1954 by Modigliani and Brumberg. Both their papers and advance copies of the permanent income theory of Milton Friedman (1957) were circulating in 1953 and led to M.R. Fisher carrying out tests of the theories even preceding publication of Friedman's work (1956). Both the Modigliani–Brumberg and the Friedman theories are referred to as life cycle theories and they certainly have many similar implications, but the one that is more closely related to the life cycle with emphasis on age – Modigliani and Brumberg – is the one to which we confine ourselves here.

The key which rendered the multi-period analysis tractable under subjective certainty was the specification that the lifetime utility function be homothetic – this permitted planned consumption for each future period to be written as a function of expected wealth as seen at the planning date, the functional parameters being in no way dependent upon wealth, but upon age and tastes. The authors further sharpened their hypothesis. They specified that an individual would plan to consume the same amount in real discounted terms each year.

Throughout, desired bequests and initial assets were set at zero. However, the authors did show how bequests could be accounted for within the homothetic utility function itself if that became necessary.

From the outset the sharp hypotheses were designed for empirical testing since, for Modigliani at least, a propelling influence had been the debate about the explanatory power of the Keynesian consumption function for forecasting postwar consumption and income. The inadequacies revealed had led already to several refined theories, notably by Duesenberry (1948) and by Modigliani himself (1949). In the 1940s cross-section studies had been carefully carried through at the NBER and empirical results therefrom were promoting theoretical insights. The names of Dorothy Brady, Rose Friedman, Margaret Reid and Janet Fisher immediately come to mind. Any new theory had to be consistent with their findings.

The tighter specification of the hypothesis enabled the spelling out of the pattern of accumulating savings in the working years to finance the retirement years – hump savings. Assuming that real income of each member of a population-wide sample remained the same throughout working life, it was shown that the marginal propensity to save in a cross-section was independent of age and the income distribution and depended only on the proportion of retirement years to expected lifetimes. This alerted economists to the fact that cross-section results do *not* directly translate into estimates of the marginal propensity to save of an individual planning function. This insight is of broader significance not confined to the simple hypothesis.

The implications of the hypothesis for time series analysis were disseminated much more slowly as the companion paper to that on cross-section interpretation was never published, accounts not being freely available until 1963 and the original text itself not until 1980. Real consumption including the depreciation of durable goods is a proportion of expected real wealth, and wealth is the addition of initial assets at planning date, current income and expected (discounted) future incomes. By then assuming that the proportionality factor referred to is identical across individuals, they devised, by addition, an aggregate relation for each and every age group. Next they proceed to aggregate across age groups. Here the proportionality factor, depending as it does on age, is not independent of assets, and bias may be introduced. If the strictest set of assumptions used in the cross-section analysis is employed, the authors show that when aggregate real income follows an exponential growth trend the parameters of the aggregate relation remain constant over time. They are, however, sensitive to the magnitude of the growth rate of real income (a sum of growth rates in productivity and population), the saving–income ratio being larger the greater the rate of income growth.

If income and/or assets at any time move out of line with previous planning expectations, plans can be revised. Suppose income rises yet income expectations are not revised, the change being viewed as a one-off event, then the individual marginal propensity to save at that date would rise to finance subsequent consumption at a higher plane until death. If income expectations were revised upwards, say in parallel, then the marginal propensity to save would also rise but to a lesser degree than in the one-off case as higher consumption can more easily be provided for out of later-period incomes. Allowance for income variability is straightforward in cross sections; with time series expected income, here labour income, may be set equal to a weighted average of aggregate past and expected future incomes, or subdivided according to whether the reference is to employed or unemployed consumers at any time (Modigliani and Ando 1963).

Testing of the Theories

Cross-Section

Modigliani and Brumberg drew on the writings of Margaret Reid, Dorothy Brady, Rose Friedman and Janet Fisher to support their propositions and used a study by Klein as the closest parallel to give assurance about the usefulness of the approach. The first independent test was by M.R. Fisher using a single-period cross-section of savings of some 2000 households. Data were subclassified by age of head of household and by socio-economic group as a proxy for income stability. Current income and liquid assets holdings were used as independent variables. There was evidence of peaking of marginal propensities to save in higher-age working groups, and rundown of assets in retirement years was evident. Negative savings were also exhibited in the youngest age groups depicted. The more variable income groups certainly seemed to save a good deal more as the theory would suggest. Fisher incorporated family size into the analysis though subsequently Modigliani suggested a somewhat tidier formulation. Fisher's results by age category were claimed by Clower to be inconsistent with both permanent and life cycle theories, yet in conformity with a modified theory in which consistency between income and wealth accounts is more carefully drawn, albeit in a less rigid hypothesis. Modigliani argued, however, that there was some doubt about the inconsistency of the results.

The tests that Modigliani thinks are critical are those that influenced the presentation of the theory itself, namely those posed by the then contemporary consumption analyses. He also expected net worth to rise with age up to retirement and fall thereafter for given levels of permanent income. For this he finds considerable support.

Doubts as to the appropriateness of the proportionality assumption have been widely expressed (Mayer 1972).

Time Series

In the absence of net worth figures, Modigliani and Ando (1963) resorted to simplifications which enabled the regression equation in aggregate

consumption to be written as a linear relation in current aggregate income and lagged consumption. This formulation has been used by others to test alternative theories and is not therefore the best discriminator for the purpose. Once Goldsmith's data on US personal wealth became available the linear equation expressing aggregates for consumption in terms of labour incomes and personal wealth was subjected to test. Results were encouraging for the US and also for the UK (Lydall, Stone, Pesaran and Evans) especially with respect to the coefficient of wealth for which the life cycle hypothesis has a clear degree of differentiation from other hypotheses. Obviously such a summary equation ignores the role of changing age distributions and their impact on income and net worth coefficients, a central theme of life cycle theory. Perhaps only cross-section data can be used to test these effectively as and when needed data become available. White (1978) tried another form of test on the aggregative function. Simulation tests were conducted premised on the foundation assumptions used by Modigliani and Brumberg and also alternative specifications that had been presented in the literature; for example, exponentially growing income, age-related income streams, size of families. In every case when considered there was a considerable (over 50%) shortfall in the predicted level of aggregate personal savings as compared with actual personal savings in the years compared. White surmises that even bequest motives cannot explain this degree of shortfall.

Critical among the independent variables of the life cycle hypothesis are expected future incomes. At first these were approximated by a distributed lag expression for past incomes updated in a growing economy by a trend factor. A theoretical development of the early 1960s, itself influenced by the permanent and life cycle theories – the rational expectations hypothesis (Muth) – induced Hall (1978) to assume that individuals know the stochastic process generating labour incomes. Then he is able to show that rational expectations enables future incomes to be replaced by a lagged consumption term and a term representing the effect of new information about changes in real incomes.

Since the hypothesis of rational expectations is subject to test, its combination with life cycle formulations may make it difficult to assign responsibility for deficiencies between the two. Hall used the permanent income hypothesis and found that stock market prices had explanatory power, something not predicted by the theory. Muellbauer (1983) and Wickens and Molana (1984) have found reason to allow for varying real interest rates though their belief in the importance of liquidity constraints has not received empirical support. Careful tests by Flavin (1981) were not very supportive of the permanent income theory with or without the Hall amendment. She did not test the life cycle theory itself though in the aggregate version the two are very closely related.

Later Developments in Theory and Practice

Emphasizing that the young and old coexist at any time, 'overlapping generations' models (of which Modigliani and Brumberg are now seen to be a special case) have been fruitful in depicting the equilibrium pattern of growth in an economy over time, in bringing into sharp relief the role of interest rates, and in weighing the welfare contributions of social security and private market savings schemes. They have also sharpened up the treatment of bequests, both anticipated and accidental (Abel 1985). They have lent themselves to simulation studies but have not proved rewarding for tests against empirical data.

Models of dynamic labour supply have been developed in a life-cycle hypothesis framework (M.R. Fisher 1971; Ghez and Becker 1975; MaCurdy 1981) and are being submitted to test as data become available.

Recent applications and extensions have related to the rapid development of social security and its effect on private savings, and variation of dates of retirement (Feldstein 1974; Kotlikoff 1984) on the one hand and effects of a switch from income or capital taxes to consumption taxes on the other (Seidman 1984). The social security studies have necessitated the use of more carefully defined wealth and income figures. All have raised questions as to the adequacy of the

life cycle model without much more attention to bequest issues or allowance for uncertainty as to date of death. In part it is argued that the life cycle may apply to a large section of the population but the big savers and even the lowest earners may obey different criteria (Kotlikoff and Summers 1981). Repeatedly in well-defined samples (e.g. Mirer 1979), though not in all, the decline in wealth with age was not significant; in more finely grouped data by cohorts it even rises with age! Most of the studies do not control for longer-run incomes. When this is done, King and Dicks-Mireaux (1982) find clear evidence of the hump-shaped profile for wealth over the life cycle in the 1977 Canadian Survey of Consumer Finances and a decline in wealth with age at a rate slower than the Modigliani model would predict. Farrell (1959) was the first to suggest the need for amendment to the theory to allow for bequests and uncertainty, but there has been no rush to incorporate his or even Modigliani's revisions. This apart, the effects of a switch from income to consumption taxes, a contemporary policy issue, lends itself extremely well to analysis in a life cycle format.

The life cycle hypothesis is a robust plant, even though a few branches are in need of pruning to make it more so.

See Also

- ▶ Consumer Expenditure
- ▶ Consumption Function
- ▶ Friedman, Milton (1912–2006)

Bibliography

Abel, A.B.. 1985. Precautionary savings and accidental bequests. *American Economic Review* 75(4): 777–791.

Ando, A., and F. Modigliani. 1957. Tests of the life cycle hypothesis of savings: Comments and suggestions. *Oxford Institute of Statistics Bulletin* 19: 99–124.

Blinder, A. 1974. *Toward an economic theory of income distribution*. Cambridge: Cambridge University Press.

Brumberg, R. 1956. An approximation to the aggregate saving function. *Economic Journal* 66: 66–72.

Clower, R.W., and M.B. Johnson. 1968. Income wealth and the theory of consumption. In *Value, capital and growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

Davies, J. 1981. Uncertain lifetime, consumption and dissaving in retirement. *Journal of Political Economy* 89(3): 561–577.

Diamond, P. 1965. National debt in a neoclassical growth model. *American Economic Review* 55: 1126–1150.

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

Farrell, M.J. 1959. The new theories of the consumption function. *Economic Journal* 69: 678–696.

Farrell, M.J. 1970. The magnitude of 'rate-of-growth' effects on aggregate savings. *Economic Journal* 80(320): 873–894.

Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82(5): 905–926.

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

Fisher, M.R. 1956. Exploration in savings behaviour. *Oxford University Institute of Statistics Bulletin* 18: 201–277.

Fisher, M.R. 1957. A reply to the critics. *Oxford University Institute of Statistics Bulletin* 19: 179–199.

Fisher, M.R. 1971. *The economic analysis of labor*. New York: St Martin's Press.

Flavin, M. 1981. The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89(5): 974–1009.

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

Ghez, G., and G.S. Becker. 1975. *The allocation of time and goods over the life cycle*. New York: Columbia University Press.

Hall, R.E. 1978. Stochastic implications of the life cycle–permanent income hypothesis: theory and evidence. *Journal of Political Economy* 86(6): 971–987.

Harrod, R. 1948. *Towards a dynamic economics*. London: Macmillan.

King, M.A., and L.D.L. Dicks-Mireaux. 1982. Asset holdings and the life cycle. *Economic Journal* 92: 247–267.

Klein, L.R. 1951. Assets, debts and economic behaviour. In *Studies in income and wealth*, vol. 14. New York: National Bureau of Economic Research.

Kotlikoff, L.J. 1984. Taxation and savings: A neoclassical perspective. *Journal of Economic Literature* 22(4): 1576–1629.

Kotlikoff, L.J., and L.H. Summers. 1981. The role of intergenerational transfers in aggregate capital accumulation. *Journal of Political Economy* 89(4): 706–732.

Landsberger, M. 1970. The life cycle hypothesis: A reinterpretation and empirical test. *American Economic Review* 60(1): 175–183.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy* 89(6): 1059–1085.

Mayer, T. 1972. *Permanent income wealth and consumption*. California: University of California Press.

Menchik, P.L., and M. David. 1983. Income distribution, lifetime savings, and bequest. *American Economic Review* 73(4): 672–690.
Mirer, T.W. 1979. The wealth–age relationship amongst the aged. *American Economic Review* 69(3): 435–443.
Modigliani, F. 1949. Fluctuations in the saving–income ratio: A problem in economic forecasting. In *Studies in income and wealth*, vol. 11. New York: National Bureau of Economic Research.
Modigliani, F. 1975. The life cycle hypothesis of saving twenty years later. In *Contemporary issues in economics*, ed. M. Parkin. Manchester: Manchester University Press.
Modigliani, F., and A. Ando. 1963. The life cycle hypothesis of saving: Aggregate implications and tests. *American Economic Review* 53: 55–84.
Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of cross-section data. In *Post-Keynesian economics*, ed. K.K. Kurihara. New Brunswick: Rutgers University Press.
Modigliani, F., and R. Brumberg. 1980. Utility analysis and aggregate consumption function. In *The collected papers of Franco Modigliani*, vol. 12, ed. A. Abel. Cambridge, MA: MIT Press.
Muellbauer, J. 1983. Surprises in the consumption function. *Economic Journal* Supplement: 34–50.
Pesaran, M.H., and R.A. Evans. 1984. Inflation capital gains and UK personal savings: 1953–1981. *Economic Journal* 94: 237–257.
Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.
Seidman, L.S. 1983. Taxes in a life cycle growth model with bequests and inheritances. *American Economic Review* 73(3): 437–441.
Seidman, L.S. 1984. Conversion to a consumption tax: The transition in a life-cycle growth model. *Journal of Political Economy* 92(2): 247–267.
Summers, L.H. 1981. Capital taxation and accumulation in a life cycle growth model. *American Economic Review* 71(4): 533–544.
Tobin, J. 1967. Life cycle saving and balanced growth. In *Ten economic studies in the tradition of Irving Fisher*, ed. W. Fellner. New York: Wiley.
White, B.B. 1978. Empirical tests of the life cycle hypothesis. *American Economic Review* 68(4): 547–560.
Wickens, M.R., and H. Molana. 1984. Stochastic life cycle theory with varying interest rates and prices. *Economic Journal* 94: 133–147.
Williamson, S.H., and W.L. Jones. 1983. Computing the impact of social security using the life cycle consumption function. *American Economic Review* 73(5): 1036–1052.
Yaari, M.E. 1965. Uncertain lifetime, life insurance and the theory of the consumer. *Review of Economic Studies* 32: 137–150.

Life Insurance

Karl H. Borch

A simple life insurance contract can be of two forms: (i) annuities paying specified amounts on fixed dates, provided that the insured is alive; or (ii) life insurances paying a specified amount at the death of the insured. All life insurance contracts can be built up as combinations of these two basic components.

The actuarial calculations in life insurance are based on the 'principle of equivalence', which requires that the expected present values of the payments made by the insured and by the insurer must be equal. If the administrative expenses of the insurer are ignored, the expected present value of his payments will be equal to that of the receipts of the insured.

The expectations are calculated from a mortality law, described by the death rate q_x, which can be interpreted as the probability that a person of age x will die before he reaches the age $x + 1$. From the death rates one calculates, usually after some graduation, the mortality table l_x by the formula

$$l_{x+1} = l_x(1 - q_x).$$

The ratio

$$_tp_x = \frac{l_{x+t}}{l_x} = \pi(t)$$

can be interpreted as the probability that a person of age x will still be alive after a time t. Table 1 gives the values of q_x under some widely used mortality tables.

If the function $\pi(t)$ is continuous it is convenient to write

$$\pi(t) = \exp\left\{-\int_0^t \mu_{x+s}\mathrm{d}s\right\},$$

where μ_x is the 'force of mortality'. The single premium for the pure endowment can then be written

Life Insurance, Table 1 Death rates: $1000q_x$ under some mortality laws

		1980 CSO	1980 CSO
x	H^{M}	Male	Female
10	4.90	0.75	0.68
30	7.72	1.75	1.37
50	15.95	7.00	5.13
70	62.19	47.37	23.16
90	279.45	228.43	198.85

H^{M}: The table from 1869 based on the experience for Healthy Males of 20 British life insurance companies. In some countries this table was used into the present century

1980 CSO: The US Commissioners' Standard Ordinary Lives tables from 1980, prepared by the National Association of Insurance Commissioners. The simplest life insurance contract is the 'pure endowment'. Under this contract a unit is paid to the insured if he is alive at time t. From the principle of equivalence it follows that the single premium for this contract is ${}_tE_x = \pi(t)e^{-\delta t}$, where δ is the 'force of interest'

$$ {}_tE_x = \pi(t)\mathrm{e}^{-\delta t} = \exp\left\{-\int_0^t \delta + \mu_{x+s})\mathrm{d}s\right\}. \quad (1) $$

From (1) it follows that the premium is less than the present value of a unit payable with certainty at time t, because the discounting is done at a higher and increasing rate of interest.

A life-long annuity is a sum of pure endowments, and hence the single premium is

$$ a_x = \sum\nolimits_{t=1}^{\infty} {}_tE_x \sum_{t=1}^{\infty} \mathrm{e}^{-\delta t}\pi(t). $$

In theoretical work it is often advantageous to assume that the annuity is paid as a continuous stream, and write

$$ \overline{a}_x = \int_0^{\infty} \mathrm{e}^{-\delta t}\pi(t)\mathrm{d}t $$

for the single premium.

Under a typical pension plan the insured will pay a constant or 'level' premium P up to the time n, and from then on he will receive an annuity B as long as he lives. The principle of equivalence gives the following relationship between premium and benefits:

$$ P = \int_0^n \mathrm{e}^{-\delta t}\pi(t)\mathrm{d}t = B\int_0^{\infty} \mathrm{e}^{-\delta t}\pi(t)\mathrm{d}t $$

or in the standard actuarial notation

$$ P\overline{a}_{x:n\rceil} = B\{\overline{a}_x - \overline{a}_{x:n\rceil}\}. \quad (2) $$

As $\pi(t)$ is non-increasing, $\pi(0) = 1$ and $\pi(\infty) = 0$, it follows that $F(t) = 1 - \pi(\mathrm{t})$ has the properties of a cumulative probability distribution. Hence $F'(t) = 1 - \pi'(t)$ can be interpreted as the probability density of the event that the person will die at time t. The present value of a unit payable then is $\mathrm{e}^{-\delta t}$. From the principle of equivalence it follows that the single premium for an insurance contract paying a unit at the time of the death of the insured is

$$ \overline{A}_x = -\int_0^{\infty} e^{-\delta t}\pi'(t)\mathrm{d}t = 1 - \delta\int_0^{\infty} \mathrm{e}^{-\delta t}\pi(t)\mathrm{d}t $$

or

$$ \overline{A}_x = 1 - \delta\overline{a}_x. $$

The continuous level premium, for the whole duration of this insurance contract is determined by

$$ P\overline{a}_x = \overline{A}_x = 1 - \delta\overline{a}_x \quad (3) $$

or

$$ P = \frac{1}{\overline{a}_x} - \delta. $$

The contract described, called whole-life insurance, remains in force until the insured dies. Two other insurance contracts in general use are:

Term insurance. The single premium is

$$ \overline{A}_{x:n\rceil} = 1 - \delta\overline{a}_{x:n\rceil} - \mathrm{e}^{-\delta n}\pi(n). $$

Under this contract the sum insured is paid only if the insured dies before the time n.

Endowment insurance with a single premium

$$\overline{A}_{x:n\rceil} = 1 - \delta\overline{a}_{x:n\rceil}.$$

Under this contract the unit sum is paid after a time n, or at earlier death. The contract is clearly the sum of a pure endowment and a term insurance. Most life insurance contracts contain an important element of saving. For the pension plan described by (2) this is obvious, and it holds also for the whole life insurance with a level premium determined by (1). At a time t after a contract of this kind was concluded, the insured will have reached the age $x + t$. He will then have accumulated savings amounting to

$$_{x}V_{t} = \overline{A}_{x+t} - P\overline{a}_{x+t}. \tag{4}$$

It can be shown that this equation also can be written in the form

$$_{x}V_{t} = \frac{e^{\delta t}}{\pi(t)}\left\{P\overline{a}_{x:\overline{t}|} - \overline{A}^{1}_{x:t}\right\}. \tag{5}$$

In (2) the first term is the expected present value, at time t, of a unit payable at the death of the insured. The second term is the expected present value of the premiums which the insured according to the contract has undertaken to pay as long as he lives. The difference between the two represents the net expected present value of the insurer's obligations under the insurance contract at time t.

The first term in braces in (3) is the expected present value at time 0, of the premiums paid by the insured up to time t, and the second term is the value of the insurance cover he has received. The difference, compounded up to time t, at the rate of interest defined by (1), gives the accumulated saving of the insured at time t.

The common left-hand side of (2) and (3) is usually called the 'premium reserve', since it represents that part of premiums received which the insurer must keep in reserve to meet his expected future obligations.

The premium reserve can also be interpreted as the accumulated savings of the insured at time t. Often he will have the right at any time to cancel the contract and receive the 'surrender value' of his policy in cash. The surrender value will usually be equal to the premium reserve, less a deduction for expenses incurred by the insurer.

An insurance contract can be of very long duration, and once the contract is concluded the insurer can usually not change the term. When he quotes a premium, the quotation must be based on forecasts of interest and mortality rates several decades into the future. It is natural, and indeed necessary, that these forecasts should include considerable safety margins. This means, however, that the insurance contract can be expected to yield a substantial 'surplus', or profit to the insurer. Under most life insurance contracts this surplus is paid back to the insured, once it has been realized.

Conventionally the theory of life insurance is formulated in terms of probabilities, although this is not really necessary. If sufficient safety margins are included in the assumptions about future interest and mortality, there is a high probability that surplus will arise. This surplus, if it materializes, is distributed to the policy-holders by methods closely related to those of cost accounting. Essentially the actual cost to the insurer is calculated for a group of similar contracts, and the excess of premiums paid over costs is refunded to the policy holders in the group.

In older insurance policies the sum payable at death was usually the same during the whole duration of the contract, and the premium was constant over the same or a shorter period. The only flexibility was that the insured could buy and surrender any combination of term and endowment contracts, as his needs changed.

A more general insurance contract consists of the two elements: (i) $C(t)$ = the amount payable at death, if death occurs at time t; (ii) $P(t)\,dt$ = the premium paid by the insured in the time interval $(t, t + dt)$. $P(t)$ may be negative for some values of t, as in a pension plan.

The principle of equivalence requires that

$$\int_0^\infty C(s)e^{-\delta s}\pi'(s)ds + \int_0^\infty P(s)e^{-\delta s}\pi(s)ds = 0. \tag{6}$$

Any pair of functions satisfying this condition represents a feasible insurance contract. In practice one will require however that the premium reserve always be non-negative.

A generalization of (3) gives the premium reserve at time t for this contract as

$$ {}_xV_t = \frac{e^{\delta t}}{\pi(t)} \int_0^t \{P(s)\pi(s) + C(s)\pi'(s)\} e^{-\delta s} ds. $$

Hence any insurance contract which satisfies (4) is possible, provided that ${}_xV_t \geq 0$ for all t, and the insured can change the contract at any time, provided that the two conditions are not violated. To avoid adverse selection there must, however, be restrictions on how the insured can increase the death benefit $C(t)$ in the contract period.

General contracts of this form have been introduced in most countries during recent decades, and they have made life insurance a very flexible instrument of saving.

The conventional insurance contract is expressed in nominal units of money, and this may lead the insurers to invest the funds in fixed interest securities. In times of inflation such investments are not very attractive, since the real rate of interest may well be negative. To meet competition from other forms of savings, insurance companies in some countries have introduced different types of 'equity-linked' insurance contracts. One way of doing this is to express the benefits to the insured in terms of units in an investment fund. This makes it possible to construct insurance contracts representing any combination of risk-free investment and the higher return associated with risky investment.

History

The history of life insurance can be traced back at least to the days of the Roman Empire. In the Middle Ages the guilds imposed special dues on their members, and the amounts collected were paid to the dependents of the members who had died during the past year. The sums involved were usually modest, and the main objective seems to have been to secure a decent funeral for the departed member.

In the 17th century mutuals or 'friendly societies' were formed in several European countries. These societies, which offered life insurance for modest amounts on principles similar to those used by the medieval guilds, often operated on a shaky technical basis. It is generally agreed (cf. Ogborn 1962) that modern life insurance began in 1762 with the formation of the Equitable Society in London. This society, which is still in operation, introduced the correct scientific methods in the calculation of its premiums and reserves.

From the 16th to the 18th century the sale of life annuities or pensions was an important element in government borrowing. Governments usually found it difficult to repay their debts on schedule, and some loans were floated without any redemption plan, such as the consols in Britain. In such cases the interest payments became a perpetual annuity. If it was agreed that payments should cease with the death of the lender, the loan would automatically be liquidated, although annual payments would be higher.

The correct formula for the expected present value of a life-long annuity was presented by Jan de Witt in 1671, in a report: *De Vardy van de Lifrenten* (The Value of Life Annuities) to the States General of the Netherlands. According to Neuburger (1974), de Witt argued that a life-long annuity to a child of 3 years should be priced at 16 times the annual payment. He was, however, overruled by the politicians, who decided that the price should be 14.

A special form of borrowing was the 'tontine', named after Lorenzo Tonti who proposed the scheme to Cardinal Mazarin. Under a tontine a large number of tickets or bonds were sold to buyers who were divided into age-groups. The government paid the agreed interest on the total amount raised by each group, and the interest was divided among the surviving members. When all members of the group had died, payments ceased, and the debt was liquidated. The tontine brought an element of gambling into the purchase of life annuities, and this seems to have been appreciated by investors at the time (cf. Jennings and Trout 1982).

The early life insurance companies were often founded by idealists who wanted to make safe life insurance available to those who wanted to provide for their old age, or for dependents in case of early death. Growth was slow, until the active selling of life insurance began in the second half of the 19th century. Since then growth has been rapid, and in the industrialized Western countries 2–4 per cent of GNP is spent on life insurance premiums.

See Also

- ▶ Insurance
- ▶ Life Tables

Bibliography

Jennings, R.M., and A.P. Trout. 1982. *The tontine: From the reign of Louis XIV to the French revolutionary era*. Homewood: Richard D. Irwin.

Neil, A. 1977. *Life contingencies*. London: William Heinemann.

Neuburger, E. 1974. Die Versicherungsmathematik von vorgestern bis heute. *Zeitschrift für die gesamte Versicherungswissenschaft* 63: 107–124.

Ogborn, M. 1962. *Equitable assurances, the story of a life assurance society*. London: Allen & Unwin.

Life Tables

Nathan Keyfitz

JEL Classifications
J16

Life tables present the age incidence of mortality in a population. The population may be all those people in a country or other area, or some category within a country; it may be all persons counted at a particular moment or period of time, say 1980 (period table); or it may be those born at a particular time and followed through life (cohort table).

The abridged life table officially calculated for the United States deaths and population of 1983 (National Center for Health Statistics 1983) is shown as Table 1. It is based on population estimated to mid-year (${}_5P_x$ for age x to $x + 4$ at last birthday) and extrapolated from the 1980 census, and corresponding deaths to residents occurring during the year 1980 (${}_5D_x$). Unregistered deaths are few in developed countries, but population censuses tend to under count, and give a life table of too high mortality unless a correction is made. In most less developed countries registration of deaths is incomplete, and model (e.g. Coale and Demeny 1983 or UN 1982) tables fill the gap.

Having the age-specific death rates, ${}_5M_x = {}_5D_x/{}_5P_x$, the important step is calculating the probability that a person living at the beginning of the age interval will survive to the end. If the death rate within the interval can be assumed constant then the exact probability is $l_{x+5}/l_x = e^{-5{}_5M_x}$. In fact, for ages from about 10 onwards the death rate rises within as well as between intervals, and this is partly taken into account by the alternative more precise expression

$$\frac{l_{x+5}}{l_x} = \frac{1 - 5{}_5M_x/2}{1 + 5{}_5M_x/2}.$$

Greville (1943) gives a more general expression. More generally yet, if $p(x + t)$ is the continuous age distribution within the interval x to $x + 5$, and $\mu(x + t)$ the continuous death rate, then we have the equation

$${}_5M_x = \frac{\int_0^5 p(x+t)\mu(x+t)\mathrm{d}t}{\int_0^5 p(x+t)\mathrm{d}t}$$

from which it is required to extract the quantity

$$l_{x+5}/l_x = \exp\left[-\int_0^5 \mu(x+t)\mathrm{d}t\right].$$

A solution (Keyfitz 1985, p. 39) is obtained by expanding the p's and the μ's by Taylor's theorem.

Having obtained the probability of surviving from one point of age to the next, the life table is

Life Tables, Table 1 Abridged life tables by race and sex: United States, 1980

Age interval	Proportion dying	Of 100,000 born alive		Stationary Population		Average remaining lifetime
Period of life between two exact ages stated in years (1)	Proportion of persons alive at beginning of age interval dying during interval (2)	Number living at beginning of age interval (3)	Number dying during age interval (4)	In the age interval (5)	In this subsequent age intervals (6)	Average number of years of life remaining of age interval (7)
X to $x + n$	${}_nq_x$	L_x	${}_nd_x$	${}_nL_x$	T_x	$\$_x$
All races						
0–1	0.0127	100,000	1266	98,901	7,371,986	73.7
1–5	0.0025	98,734	250	394,355	7,273,085	73.7
5–10	0.0015	98,484	150	492,017	6,878,730	69.8
10–15	0.0015	98,334	152	491,349	6,386,713	64.9
15–20	0.0049	98,182	482	489,817	5,895,364	60.0
20–25	0.0066	97,700	648	486,901	5,405,547	55.3
25–30	0.0066	97,052	638	483,665	4,918,646	50.7
30–35	0.0070	96,414	672	480,463	4,434,981	46.0
35–40	0.0091	95,742	875	476,663	3,954,518	41.3
40–45	0.0139	94,867	1321	471,250	3,477,855	36.7
45–50	0.0222	93,546	2079	462,857	3,006,605	32.1
50–55	0.0351	91,467	3209	449,811	2,543,748	27.8
55–60	0.0530	88,258	4676	430,230	2,093,937	23.7
60–65	0.0794	83,582	6638	402,081	1,663,707	19.9
65–70	0.1165	76,944	8965	363,181	1,261,626	16.4
70–75	0.1694	67,979	11,517	312,015	898,445	13.2
75–80	0.2427	56,462	13,702	248,534	586,430	10.4
80–85	0.3554	42,760	15,197	175,192	337,896	7.9
85 and over	1.0000	27,563	27,563	162,704	162,704	5.9

Source: National Center for Health Statistics (1983)

completed by cumulating these probabilities from age 0; with an arbitrary starting point ('radix') of 1 or 100,000 the l_x column is obtained by successive multiplication

$$l_{x+5} = l_x\left(\frac{l_{x+5}}{l_x}\right), \quad \text{etc.}$$

The l_x column has three interpretations: (1) The probability of a person just born surviving to age x. (2) The number of survivors in a hypothetic cohort (say starting with 100,000 births) by the time age x is reached. (3) The number of persons aged x in the stationary population.

For this last interpretation one integrates over one- or five-year age intervals, and so obtains ${}_5L_x = \int_0^5 l(x+t)\mathrm{d}t$, the number of individuals in a stationary population (say one in which there are exactly 100,000 births per year) at age x to $x + 4$ at last birthday.

What makes possible the simultaneous representation of these three quite different entities is a central assumption of the life table: that the actual number of deaths to occur will be the probability multiplied by the initial number exposed. In short, the life table is a deterministic model: if there are a million people, each with a probability of 0.01 of dying during the following year, there will be exactly 10,000 deaths. It also assumes that every individual of a given age and sex has the same probability of dying.

The estimators above do not make explicit allowance for withdrawals, nor for the individual

times at death. With small populations, for instance those used in follow-up studies after a diagnosis of cancer, or after a particular treatment, more refined methods are needed. One such, called the product-limit method and using maximum likelihood, is due to Kaplan and Meier (1958). This and ways of dealing with withdrawals and censoring are taken up in Elandt-Johnson and Johnson (1980, ch. 6).

From the probability of surviving the life expectancy is calculated as $e_x^0 = \int_0^{\omega - x} l(x+t)\mathrm{d}t$. In the deterministic model this traces a (usually synthetic) cohort consisting of l_x individuals, who will live ${}_5L_x$ person-years over the next five years; ${}_5L_x + 5$ over the five years after that, etc. These future years may be thought of as divided among the l_x persons, giving each of them an average of $e_x^0 = \sum L_{x+h}/l_x$.

An original purpose of the life table was to calculate annuities and life insurance, and much of the modern notation has been developed by actuaries. If money carried no interest then the value of an annuity starting at age 65, say, would be $\int_{65}^{\omega} l(t)\mathrm{d}t$, and if this was to be paid for by yearly payments from age 20, the annual premium would be $\int_{65}^{\omega} l(t)\mathrm{d}t / \int_{20}^{65} l(t)\mathrm{d}t$. If money carries interest we need to discount this (say back to birth), and the premium will be less, being calculated as $\int_{65}^{\omega} e^{-it} l(t)\mathrm{d}t / \int_{20}^{65} e^{-it} l(t)\mathrm{d}t$, where i is the rate of interest compounded momently (Jordan 1967).

The expectation of life at age 0 is a common measure of mortality, for comparing countries and other population aggregates: in the United States the life expectancy was 75 years in 1983, compared with 66 years for Mexico. Mexico's crude rate (1000 × D/P) is 6 per thousand against the US 9, a comparison that does not reflect true mortality because Mexico's population is much younger.

The third meaning of the life table can be generalized to represent the age distribution of a population that is increasing at a steady rate r; in this generalization the number of persons aged x to $x + 4$ at last birthday is proportional to $\int_0^5 e^{-(x+rt)} l(x+t)\mathrm{d}t$.

The life table idea is readily extended to more than two states of exit. One can work out the chance of dying from the several possible causes of death – cancer, heart disease, etc.; this is still a decrement table, but now with several causes of decrement.

While the notation and the concepts of the life tables were worked out for mortality, it is applied to many processes other than living and dying. A woman has a certain probability month by month of becoming pregnant; the probabilities can be cumulated to give the probability of still not being pregnant by the xth month, from which the expected months to pregnancy can be calculated for women who are fertile. An aircraft engine has a certain probability of breaking down in the first month, the second month, etc.; a life table shows the expected number of months of service, and by an extension the number of engines that will have to be kept in reserve for replacements up to a given level of security. Biological ecologists calculate life tables for many species of animals and insects. Probability of divorce in the first year, the second year, etc. after marriage can be worked out in the same two- state model, except that now first marriage and divorce rather than living and dead are the states in question. A table can be made for survival within the school system, in which the states are attending school and dropping out.

In increment–decrement tables persons can re-enter some of the states. For instance they can enter the labour force, then leave it, then enter again. The same applies to marriage, or to migration among regions of a country. For this a multidimensional analogue of the life table is available, and has been extensively used (Rogers 1975, 1984; Schoen 1975). The relevant formulas are matrix analogues of the ordinary life table formulas given above.

A main use of life tables is for population projection. If the population age x to $x + 4$ at the jumping-off point is ${}_5P_x$, then 5 years later it will be ${}_5P_x{}_5L_x + {}_5/{}_5L_x$ if the life table is appropriate and random variation and migration can be disregarded. (For the birth component and other aspects of projection, see Brass 1974.)

In pursuing these and other purposes, one often deals with populations for which mortality data are deficient or altogether lacking. A common

procedure in the past was to substitute a suitable member of a series (for example, England and Wales at an appropriate date). Today it is more convenient to use one of the sets of model tables calculated for the purpose, based not on one country, but on all the countries for which reliable data are available (UN 1982; Coale and Demeny 1983).

Life tables are calculated on the (unrealistic) assumption that the population is homogeneous in respect of all unmeasured variables. Because the observed population is constantly being selected towards persons of greater robustness, the true expectation for a person initially of average robustness is less (by something of the order of one year) than that shown by published tables (Vaupel and Yashin 1985).

See Also

- ▶ Economic Demography
- ▶ Fertility in Developing Countries
- ▶ Graunt, John (1620–1674)
- ▶ Historical Demography
- ▶ Mortality
- ▶ Stable Population Theory

Bibliography

Brass, W. 1971. On the scale of mortality. In *Biological aspects of demography*, ed. W. Brass, 69–110. London: Taylor & Francis.

Brass, W. 1974. Perspectives in population prediction, illustrated by the statistics of England and Wales. *Journal of the Royal Statistical Society, Series A* 137: 532–583.

Chiang, C.L. 1984. *The life table and its applications*. Malabar: Robert E. Krieger.

Coale, A.J. 1984. Life table construction on the basis of two enumerations of a closed population. *Population Index* 50 (2): 193–213.

Coale, A.J., and P. Demeny. 1983. *Regional model life tables and stable populations*. 2nd ed. New York: Academic.

Elandt-Johnson, R.C., and N.L. Johnson. 1980. *Survival models and data analysis*. New York: Wiley.

Greville, T.N.E. 1943. Short method of constructing abridged life tables. *Record American Institute Actuaries* 32: 29–43.

Jordan, C.W. 1967. *Life contingencies*. Chicago: Society of Actuaries.

Kaplan, E.L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481.

Keyfitz, N. 1985. *Applied mathematical demography*. 2nd ed. New York: Springer.

National Center for Health Statistics. 1983. *Advance report of final mortality statistics, 1980*, vol. 32, no. 4, Supplement, US Department of Health and Human Services.

Rogers, A. 1975. *Introduction to multiregional mathematical demography*. New York: Wiley.

Rogers, A. 1984. *Migration, urbanization, and spatial population dynamics*. Boulder: Westview Press.

Schoen, R. 1975. Constructing increment-decrement life tables. *Demography* 12: 313–324.

United Nations. 1982. *Model life tables for developing countries*. New York: United Nations.

Vaupel, J.W., and A.I. Yashin. 1985. The deviant dynamics of death in heterogeneous populations. In *Sociological methodology 1985*, ed. N.B. Tuma. San Francisco: Jossey-Bass.

Likelihood

A. W. F. Edwards

A statistical model for phenomena in the sciences or social sciences is a mathematical construct which associates a *probability* with each of the possible outcomes. When two different models are to be compared as explanations for the same observed outcome, or perhaps two variants of the same model differing only in the value of some adjustable parameter, the probability of obtaining this particular outcome can be calculated for each, and is then known as the *likelihood* for the hypothesis (or parameter value) given the particular outcome or 'data'.

Likelihoods and probabilities are easily (and frequently) confused, and it is for this reason that in 1921 R.A. Fisher introduced the separate word 'likelihood' to draw attention to the different properties and uses of the two concepts.

The first of these is that the variable quantity in a likelihood statement is the hypothesis, the outcome being that actually observed, in contrast to the usual form of a probability statement which refers to a variety of outcomes, the hypothesis

being assumed, or fixed. Thus a manufacturer of aircraft components using a well-tried process giving a known proportion of defective items will be able to calculate the *probabilities* with which 1, 2, 3,... defective components will appear in a batch, and will be able to plan his inspection procedures accordingly; but when he later changes to an improved manufacturing process with an as-yet unknown proportion of defectives he will be able to calculate the *likelihoods* of various proportions given the numbers of defective items actually observed in a particular batch. As we shall see, such likelihoods provide information about the true, but unknown, proportion.

The second different property arises directly from the first. If all the outcomes of a statistical model are considered, their total probability will be 1 since one of them must occur and they are mutually exclusive; but since, in general, hypotheses are not exhaustive – one can usually think of another one – it is not to be expected that the sum of two or more likelihoods has any particular meaning, and indeed there is no addition law for likelihoods corresponding to the addition law for probabilities. It follows that it is only *relative* likelihoods that are informative; absolute values are not relevant.

The most important application of likelihood is in parametric statistical models. To take Fisher's original example (1921), the distribution of the sample correlation coefficient r depends, in the case of the bivariate normal model, only on the value of the correlation parameter ρ of the model. Thus for any assumed value of ρ, the distribution of r for samples of a given size may be computed. But 'What we can find from a sample is the *likelihood* of any particular value of ρ, if we define the likelihood as a quantity proportional to the probability that, from a population having that particular value of ρ, a sample having the observed value r should be obtained. So defined, probability and likelihood are quantities of an entirely different nature.'

By way of notation, let $P(R/\rho)$ be the probability density function of the random variable R given the population parameter ρ. Then we write

$$L(\rho \| r) \propto \mathrm{P}(r|\rho)$$

for the likelihood of ρ given a particular value r, the double vertical line $\|$ being used to indicate that the likelihood of ρ is not *conditional on r* in the technical probability sense. In the example of the correlation coefficient $L(\rho\|r)$ is a continuous function of $\rho(-1 \leq \rho \leq +1)$, known as the *likelihood function*.

The value of ρ which maximizes $L(\rho\|r)$ for an observed r is known as the *maximum-likelihood estimate* of ρ and is denoted by $\widehat{\rho}$; expressed in general form as a function of r it is known as the *maximum-likelihood estimator*. Since the pioneering work of Fisher in the early 20th century it has been known that maximum-likelihood estimators possess certain desirable properties under repeated-sampling (consistency and asymptotic efficiency), and for this reason they have come to occupy a central position in repeated-sampling theories of statistical inference.

However, partly as a reaction to some unsatisfactory features which repeated-sampling theories display, and partly as a defence against a full-blown Bayesian theory of statistical inference, likelihood has been increasingly seen as a fundamental concept enabling hypotheses and parameter values to be compared directly.

The basic notion, introduced by Fisher in 1912 whilst still an undergraduate at Cambridge, is that the likelihood ratio between two hypotheses or parameter values is to be interpreted as the degree to which the data support the one hypothesis against the other. Thus a likelihood ratio of 1 corresponds to indifference between the hypotheses, whereas the maximum-likelihood value of a parameter is the value best-supported by the data, other values being ranked by their lesser likelihoods accordingly.

Such an approach, unsupported by any appeals to repeated-sampling criteria, is ultimately dependent on the primitive notion that, other things being equal, the best hypothesis or parameter-value is the one which would explain what has in fact happened with the highest probability. The strong intuitive appeal of this can be captured by recognizing that it is the value which would lead,

on repeated sampling, to a precise repeat of the observed data with the least expected delay. In this sense it offers the best statistical explanation of the data.

In addition to specifying that relative likelihoods measure degrees of support, the likelihood approach requires us to accept that the likelihood contains all the information we can extract from the data about the hypotheses in question on the assumption of the specified statistical model – the so-called *Likelihood Principle*. These two ideas are conveniently expressed together as:

The Likelihood Axiom

Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other (Edwards 1972).

The likelihood approach has many advantages apart from its intuitive appeal. It is easy to apply because the likelihood function is usually simple to obtain analytically or easy to compute numerically. It leads directly to the important statistical concept of sufficiency, and illuminates many of the controversies surrounding repeated-sampling theories of inference, especially those concerned with ancillarity and conditioning. Likelihoods are multiplicative over independent data sets, facilitating the combination of information (for this reason *log-likelihoods*, or *supports*, are often preferred because information is then combined by addition). Most importantly it is compatible with Bayesian statistical inference in that the posterior Bayes distribution for a parameter is, by Bayes' Theorem, found by multiplying the prior distribution by the likelihood function. Thus, where a parameter *distribution* can be countenanced (and this is the Achilles' heel of Bayesian inference) all the information the data contain about the parameter is transmitted via the likelihood function, in accordance with the Likelihood Principle. It is indeed difficult to see why the medium through which such information is conveyed should depend on the purely external question of whether the parameter may be considered to have a probability distribution, and this is another powerful argument in favour of the Likelihood Principle.

There are disadvantages to the likelihood approach, however, though some of these may be attributed to its relatively undeveloped state. It is not always clear how to extract information about a parameter of interest in the presence of other unknown parameters (so-called 'nuisance parameters'), and the comparison of likelihoods for hypotheses with differing degrees of freedom is problematical. This last difficulty is probably associated with the lack of any notion of 'goodness-of-fit' in the likelihood approach, and future work may well remedy this by admitting the need for the incorporation of some repeated-sampling ideas of goodness-of-fit.

In practical terms the adoption of the Likelihood Axiom as the basis of statistical inference often means little more than a re-interpretation of existing practices, since maximum-likelihood estimates are already so widely used, but in terms of theory it brings a great clarification to large areas of statistics, sweeping away many problems associated with, for example, the interpretation of confidence intervals.

Widely-used already in the biological sciences, especially genetics, likelihood is a powerful notion wherever statistical *inference* is required; however, it is not relevant to decision theory and may therefore be expected to have a lesser impact in fields where action, rather than pure inference, is the goal.

In spite of having been widely discussed by statisticians interested in the logic of inference throughout most of the 20th century, the only book devoted exclusively to it is *Likelihood* (Edwards 1972, 1984), which contains comprehensive references to earlier work, especially that of R.A. Fisher; of conventional statistical textbooks only that by Cox and Hinkley (1974) contains relevant material. The history of likelihood is given by Edwards (1974).

See Also

- ▶ Fisher, Ronald Aylmer (1890–1962)
- ▶ Maximum Likelihood
- ▶ Probability
- ▶ Statistical Inference
- ▶ Subjective Probability

Bibliography

Cox, D.R., and D.V. Hinkley. 1974. *Theoretical statistics*. London: Chapman & Hall.
Edwards, A.W.F. 1972. *Likelihood*. Cambridge: University Press. Paperback edn, 1984.
Edwards, A.W.F. 1974. The history of likelihood. *International Statistical Review* 42: 9–15.
Fisher, R.A. 1912. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* 41: 155–160.
Fisher, R.A. 1921. On the 'Probable Error' of a coefficient of correlation deduced from a small sample. *Metron* 1(4): 3–32.

Limit Pricing

Stephen Martin

Abstract

The central idea of limit pricing is that an incumbent monopolist or collusive group will or can forestall entry by charging some price below that which maximizes static own profit. But a strategic price response is only one possible incumbent response to entry. A full understanding of the determinants of equilibrium market structure in inherently oligopolistic industries must take the full range of possible responses into account.

Keywords

Advertising; Commitment; Entry; Excess capacity; Limit pricing; Potential competition; Predatory pricing

JEL Classifications

D4

Modern economists generally trace models of limit pricing to Modigliani (1958). The idea of limit pricing is closely related to, and often not distinguished from, the much older idea that potential competition will induce a profit-maximizing incumbent monopolist or dominant group to set a price that would allow it (or, in some formulations, an entrant) only a normal rate of return (Giddings 1887; Gunton 1888; Liefmann 1915; Marshall 1890, p. 270; 1919, pp. 397–8, 524; Kaldor 1935). With this second idea, it is the presence of potential entrants that constrains the options of incumbents, not the other way around.

Modigliani's (1958) more-than-a-book-review of Bain (1956) and Sylos-Labini (1957) offered a formal model based on what Modigliani called the Sylos postulate (1958, p. 217) 'that potential entrants behave as though they expected existing firms to adopt the policy . . . of maintaining output' in the face of entry. Given such beliefs, if incumbents produce a sufficiently large output that the best post-entry price a profit-maximizing entrant could expect would be below its average cost, entry would not occur.

Gaskins (1971) generalizes the static limit price model to a dynamic context, with a model in which incumbent pricing determines the rate of expansion of a fringe of price-taking suppliers. This might also be regarded as a dynamic generalization of the familiar Forchheimer–Auspitz–Lieben model of a dominant firm in a market with a price-taking fringe.

Friedman (1979) points out that, under conditions of complete and perfect information, profit-maximizing incumbents would not, in general, maintain post–entry output at pre–entry levels, and entrants would not expect them to do so. Much the same point had been made, less formally, by Bain (1949, p. 452). Without commitment, a low price fails as an entry-limiting device if entrants believe that in the post-entry market incumbents will act in their own self-interest.

One line of research that seeks to finesse the unsatisfactory nature of the Modigliani–Sylos postulate can be traced to Spence (1977) and Dixit (1979). They offer models in which an incumbent's pre–entry investment (in capacity) alters the incumbent's post-entry incentives, and by so

doing gives credibility to post–entry conduct that renders entry unprofitable. See Allen et al. (2000) for careful discussion. The vast literature on strategic entry deterrence (Salop and Scheffman 1983; Fudenberg and Tirole 1984) springs from this root.

An alternative approach is taken by Kreps and Wilson (1982) and Milgrom and Roberts (1982). They give up the assumption of complete information and model entry-limiting behaviour based on an incumbent firm's reputation or entrants' uncertainty about an incumbent's costs. The modelling techniques employed here have been generalized to analyse predation and the conduct of regulation/competition policy under conditions of uncertainty.

The development of internally consistent theoretical models in which entry- limiting behaviour might occur as an equilibrium phenomenon was a major step in laying the game-theoretic foundation for modern industrial economics. The assembling of empirical evidence on the occurrence of limit pricing and other strategic reactions to entry has similarly followed the general trend of empirical research in industrial economics, studying particular markets for specific instances of entry-deterring behaviour.

There are case studies of limit-pricing behaviour (Blackstone 1972). But empirical studies of entry suggest that theoretical models of entry and entry deterrence abstract from essential aspects of the phenomena (Simon 2005, p. 1230).

Some real-world entry, no doubt, is like the entry of limit price and other entry-deterrence models – entry at large-scale into production of a standardized product hitherto offered by a small number of firms themselves aware of their oligopolistic interdependence. Archer Daniel Midland's well-known 1991 entry into lysine production is a case in point. Much more often, however, entry seems like the act of beginning small-scale production at a point on a Hotelling line, when location in characteristic space is largely fixed after entry and neither the entrant nor incumbents have a terribly good idea of the distribution of consumers in the region near the entrant's location.

Geroski (1995, p. 433) concludes in his careful survey that 'price is not frequently used by incumbents to deter entry, but that marketing activities are', and (1995, p. 434, fn. 7) 'work that has tried to test for the presence of limit pricing ...in general . . . has produced somewhat ambiguous results. . .. Studies of the strategic use of excess capacity to block entry have also generally produced weak and fairly unpersuasive evidence on its importance.'

Empirical work suggests that the incumbent response to entry will vary with entrant and incumbent characteristics. The response to entry will sometimes be by lowering price, sometimes by other rival strategies, and sometimes by accommodating entry.

Scott Morton (1997) finds that longer-established entrants into turn-of-the-19th-century shipping cartels were less likely to evoke a hostile response, as were entrants with substantial financial resources. Podolny and Scott Morton (1999) suggest that a predatory response was less likely if social factors were present that would allowincumbents and entrants to judge each others' 'types'. Thomas (1999) finds that incumbent US breakfast cereal manufacturers are more likely to respond with advertising to entry into a product group by other incumbents, and more likely to lower price in response to entry by a new firm. Yamawaki (2002) finds a price response to Japanese entry by German manufacturers of luxury cars for the US market, but no such response by British manufacturers. In a study of entry into the US magazine industry, Simon (2005) finds that multi-market and single-market incumbents respond differently to entry. Multi-market incumbents are more likely to cut price in response to entry by a new firm, single-market incumbents more likely to cut price in response to entry by an established publisher. He also finds that a hostile response to entry is more likely the more concentrated the target market. Conlin and Kadiyali (2006) find some evidence of the use of excess capacity as an entry-deterring device in the Texas hotel market, and also that the maintenance of excess capacity is more likely by larger firms and by firms in more concentrated markets.

It thus appears that, while a strategic price response is one possible incumbent response to entry, it is only one. A full understanding of the determinants of equilibrium market structure in inherently oligopolistic industries must take the full range of possible responses into account.

See Also

- ▶ Barriers to Entry
- ▶ Market Structure
- ▶ Monopoly
- ▶ Oligopoly
- ▶ Predatory Pricing

Bibliography

Allen, B., R. Deneckere, T. Faith, and D. Kovenock. 2000. Capacity precommitment as a barrier to entry: A Bertrand-Edgeworth approach. *Economic Theory* 15: 501–530.

Auspitz, R., and R. Lieben. 1889. *Untersuchen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

Bain, J.S. 1949. A note on pricing in monopoly and oligopoly. *American Economic Review* 39: 448–464.

Bain, J.S. 1954. Conditions of entry and the emergence of monopoly. In *Monopoly and competition and their regulation*, ed. E.H. Chamberlin. London: Macmillan.

Bain, J.S. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Blackstone, E.A. 1972. Limit pricing and entry in the copying machine market. *Quarterly Review of Economics and Business* 12: 57–65.

Conlin, M., and V. Kadiyali. 2006. Entry-deterring capacity in the Texas lodging industry. *Journal of Economics and Management Strategy* 15: 167–185.

Dixit, A. 1979. A model of duopoly suggesting a theory of entry barriers. *Bell Journal of Economics* 10: 20–32.

Forchheimer, K. 1908. Theoretisches zum unvollstandigen Monopole. In *Schmoller's Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft*, vol. 32. Munich/Leipzig: Duncker & Humblot.

Friedman, J.W. 1979. On entry preventing behavior and limit price models of entry. In *Applied game theory*, ed. S.J. Brams and G. Schwodiauer. Wurzburg/Vienna: Physica-Verlag.

Fudenberg, D., and J. Tirole. 1984. The fat-cat effect, the puppy-dog ploy, and the lean and hungry look. *American Economic Review* 74: 361–366.

Gaskins, D.W. Jr. 1971. Dynamic limit pricing: Optimal limit pricing under threat of entry. *Journal of Economic Theory* 3: 306–322.

Geroski, P.A. 1995. What do we know about entry? *International Journal of Industrial Organization* 13: 421–440.

Giddings, F.H. 1887. The persistence of competition. *Political Science Quarterly* 2: 62–78.

Gunton, G. 1888. The economic and social aspect of trusts. *Political Science Quarterly* 3: 385–408.

Kaldor, N. 1935. Market imperfection and excess capacity. *Economica* 2: 33–50.

Kreps, D.M., and R. Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27: 253–279.

Liefmann, R.L. 1915. Monopoly or competition as the basis of a government trust policy. *Quarterly Journal of Economics* 29: 308–325.

Marshall, A. 1890. Some aspects of competition. Presidential address to the economic science and statistics section of the British Association, Leeds. In *Memorials ofAlfred Marshall*, ed. A.C. Pigou. London: Macmillan. 1925.

Marshall, A. 1919. *Industry and trade*. 4th ed. London: Macmillan. 1923.

Milgrom, P., and J. Roberts. 1982. Limit pricing and entry under incomplete information: An equilibrium analysis. *Econometrica* 50: 443–466.

Modigliani, F. 1958. New developments on the oligopoly front. *Journal of Political Economy* 66: 215–232.

Podolny, J.M., and F.M. Scott Morton. 1999. Social status, entry and predation: The case of British shipping cartels 1879–1929. *Journal of Industrial Economics* 47: 41–67.

Salop, S.C., and D.T. Scheffman. 1983. Raising rivals' costs. *American Economic Review* 73: 267–271.

Scott Morton, F. 1997. Entry and predation: British shipping cartels 1879–1929. *Journal of Economics and Management Strategy* 6: 679–724.

Simon, D. 2005. Incumbent pricing responses to entry. *Strategic Management Journal* 26: 1229–1248.

Spence, A.M. 1977. Entry, capacity, investment oligopolistic pricing. *Bell Journal of Economics* 8: 534–544.

Sylos-Labini, P. 1957. *Oligopolio e Progresso Tecnico*. Milano: Giuffrè.

Sylos-Labini, P. 1962. *Oligopoly and technical progress*. Cambridge, MA: Harvard University Press.

Thomas, L.A. 1999. Incumbent firms' response to entry: Price, advertising, and new product introduction. *International Journal of Industrial Organization* 17: 527–555.

Yamawaki, H. 2002. Price reactions to new competition: A study of US luxury car market, 1986–1997. *International Journal of Industrial Organization* 20: 19–39.

Limited Dependent Variables

Takeshi Amemiya

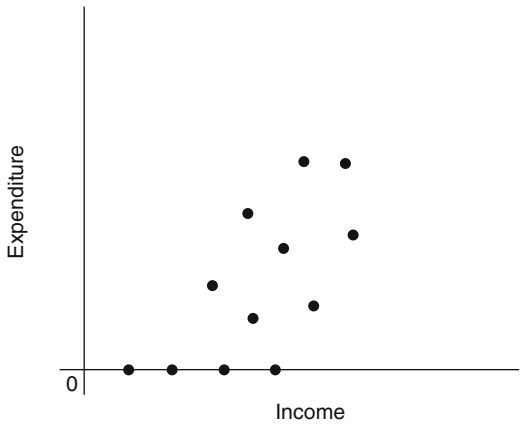
Introduction

The term *limited dependent variable* was first used by Tobin (1958) to denote the dependent variable in a regression equation that is constrained to be non-negative. In Tobin's study

the dependent variable is the household's monetary expenditure on a durable good, which of course must be non-negative. Many other economic variables are non-negative. However, non-negativity alone does not invalidate standard linear regression analysis. It is the presence of many observations at zero which causes bias to the least squares estimator and requires special analysis. For example, Tobin's data contain many households for which the expenditure on a durable good in a given year is zero.

Figure 1 shows a scatter diagram of a hypothetical expenditure–income relationship. It is clear from the diagram that the linear least squares fit of all the points will not accurately describe the relationship between expenditure and income. Later I shall indicate what statistical model will generate the data such as depicted in Fig. 1 and what estimators are more appropriate than least squares.

Tobin's model may be called a *censored regression model*. The word *censored* in this context refers to a situation where a researcher knows both the number of observations for which the dependent variable takes zero value and the value of the independent variables for those observations. In contrast, in the *truncated* regression model, those zero observations are totally lost for a researcher. An example of the data for a truncated regression model is obtained by removing the four dots lying on the horizontal axis in Fig. 1.



Limited Dependent Variables, Fig. 1 An example of a non-negative dependent variable

The study of censored or truncated data in the independent and identically distributed (i.i.d.) case predates Tobin's study in the statistical literature. However, Tobin was the first to generalize the analysis to a regression model. Censored or truncated regression models are now extensively used in many disciplines, including economics.

Many generalizations of Tobin's model have been proposed, a very simple example being to constrain the dependent variable to lie in an interval, not necessarily $[0, \infty]$. A more interesting generalization is to consider a model which involves more than one limited dependent variable. In this entry I shall discuss the three most frequently used multivariate limited dependent models: Gronau's wage-rate model, Heckman's labour supply model, and the endogenous switching regression model.

For greater detail than is possible here, the reader is referred to Maddala (1983) and Amemiya (1985). *Discrete choice models*, which are closely related to censored regression models, are discussed elsewhere in this Dictionary.

L

Tobin's Model

Definition

Tobin's model for explaining a household's expenditure on a durable good can be derived from a simple theory of utility maximization subject to the budget constraint and a boundary constraint. Define the following symbols:

y a household's expenditure on a durable good.
y_0 the price of the cheapest available durable good.
z all the other expenditures.
x income.

A household is assumed to maximize utility $U(y, z)$ subject to the budget constraint $y + z \leq x$ and the boundary constraint $y \geq y_0$ or $y = 0$. Suppose y^* is the solution of the maximization subject to $y + z \leq x$ but not subject to the other

constraint, and assume $y^* = \beta_1 + \beta_2 x + u$, where u may be interpreted as the sum of all the unobservable variables which affect the utility function. Then the solution to the original problem, denoted by y, is defined by

$$y = y^* \text{if } y^* > y_0 = 0 \text{ or } y_0 \text{ if } y^* \le y_0 \quad (2.1)$$

Now, if we assume further that u is i.i.d. over individual households with a normal distribution and that y_0 is the same for all the individual households, we obtain the following statistical model:

$$\begin{aligned} y_i^* &= x_i'\beta + u_i \\ y_i &= y_i^* \text{ if } y_i^* > 0 = 0 \quad \text{if} \quad y_i^* \le 0, \quad i = 1, 2, \ldots, n, \end{aligned} \quad (2.2)$$

where u_i are independent and identically distributed as $N(0, \sigma^2)$. This is the model proposed and estimated by Tobin (1958). It is sometimes called the *Tobit* model in analogy to the probit model. If we call its various generalizations also by the name of Tobit models, (2.2) may be called the *standard Tobit* model.

The statistical model (2.2) will produce data like those shown in Fig. 1. I shall consider various estimators of the parameters β and σ^2 in this model in the next two subsections.

Estimation

Earlier I noted that the least squares method applied to all the observations of Fig. 1 will yield biased estimates. This can be mathematically demonstrated for model (2.2) as follows. From (2.2) we obtain

$$\begin{aligned} Ey_i &= P(x_i'\beta + u_i > 0) \times E(x_i'\beta + u_i | x_i'\beta + u_i > 0) \\ &\Phi(x_i'\alpha) \times [x_i'\beta + \sigma\lambda(x_i'\alpha)] \end{aligned} \quad (2.3)$$

where $\alpha = \beta/\sigma, \lambda(x_i'\alpha) = \phi(x_i'\alpha)/\Phi(x_i'\alpha)$, and Φ and φ are the standard normal distribution and density function, respectively. Thus, the least squares estimator is biased to the extent that the last expression of (2.3) is not equal to $x_i'\beta$.

The least squares applied to only the positive observations also produces bias, although this is not apparent from Fig. 1. This can be mathematically demonstrated by considering.

$$\begin{aligned} E(y_i | y_i > 0) &= E(x_i'\beta + u_i | x_i'\beta + u_i > 0) \\ &= x_i'\beta + \sigma\lambda(x_i'\alpha), \end{aligned} \quad (2.4)$$

which is clearly not equal to $x_i'\beta$.

The term $\lambda(x_i'\alpha)$ which appears in both (2.3) and (2.4) is called *Mill's ratio* and plays an important role in a simple consistent estimator to be discussed later.

The least squares estimator, whether it is applied to all the observations or to only the positive observations, is not only biased but also inconsistent. A consistent and asymptotically efficient estimator is provided, as usual, by the maximum likelihood (ML) estimator. Tobin (1958) used it in the empirical work reported in his article. The likelihood function of Tobin's model (2.2) is given by

$$L = \prod_0 [1 - \Phi(x_i'\alpha)] \prod_1 \sigma^{-1} \phi[(y_i - x_i'\beta_i)/\sigma], \quad (2.5)$$

where Π_0 refers to the product over those i for which $y_i = 0$, and Π_1 for $y_i > 0$. The first term is equal to the probability of the observed event $x_i'\beta + u_i < 0$ and the second term is equal to the density of the observed y_i,. Thus, the likelihood function is the product of probabilities and densities. Despite this unusual characteristic of the likelihood function, it can be shown that the ML estimator is consistent and asymptotically normal with its asymptotic variance–covariance matrix given by the usual formula $-[E\partial^2 \log L/\partial\theta\partial\theta']^{-1}$. See Amemiya (1973).

Note that in Tobin's model (2.2), we observe $x_i'\beta + u_i$ when it is positive. If, instead, we do not observe it and merely learn that $x_i'\beta + u_i$ is positive, we have the so-called probit model. The likelihood function of the probit model is given by

$$L = \prod_0 [1 - \Phi(x_i'\alpha)] \prod_1 \Phi(x_i'\alpha). \quad (2.6)$$

The probit ML estimator of α, which maximizes the above, is consistent but not as asymptotically efficient as the ML estimator which maximizes (2.5). Moreover, as one can see from the form of the likelihood function (2.6), one cannot estimate β and σ^2 separately by the probit ML estimator.

Heckman (1976) noted that by inserting the probit ML estimator of α into the right-hand side of (2.4), one can obtain consistent estimates of β and σ by least squares, that is, by regressing positive y_i's on x_i and $\lambda(x_i'\hat{\alpha})$, where $\hat{\alpha}$ is the probit ML estimate. This estimator is called *Heckman's two-step estimator* and can also be used for generalized Tobit models, which I shall discuss later. In fact, it is more useful for those models than for the standard Tobit model because the ML estimator is computationally burdensome for some of the generalized Tobit models. Heckman used his estimator in the two-equation Tobit model, to be discussed in section "Heckman's Labour Supply Model" below.

Heckman's principle can be similarly applied to equation (2.3). In that case, one would regress all the y_i's, both positive and zero on $\Phi(x_i'\hat{\alpha})x_i$ and $\Phi(x_i\hat{\alpha})$. Not much is known as to which method is more preferred.

Several Monte Carlo studies have shown that Heckman's estimator can be considerably less efficient than the ML estimator in certain cases.

Nonstandard Conditions

It is well known that the least squares estimator (or equivalently, the ML estimator) in the standard normal regression model retains its consistency (although not its efficiency) when some of the basic assumptions of the model – namely, normality, homoscedasticity, and serial independence – are removed. In contrast, it has been shown that the ML estimator derived under the assumptions of model (2.2) is no longer consistent when u_i is not normal or when u_i is not homoscedastic, although it is consistent if the u_i are serially correlated. The same is true of the probit ML estimator and Heckman's estimator.

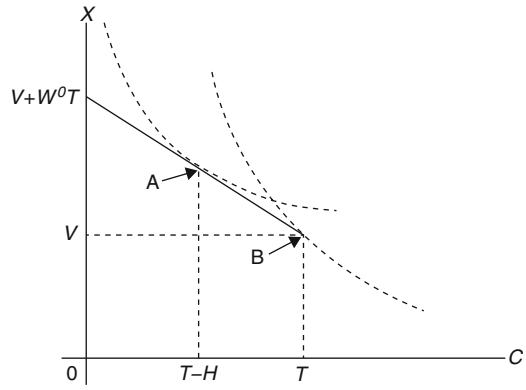
This is a serious problem because non-normality and heteroscedasticity are common occurrences in econometrics. It is recommended, therefore, that a researcher should perform a statistical test against non-normality or heteroscedasticity, and if their intensity is suspected to be high, one should incorporate non-normality or heteroscedasticy directly into one's model.

I conjecture that the normal ML estimator will do reasonably well if the degree of non-normality or heteroscedasticity is 'small'. A Monte Carlo study has shown that the normal Tobit ML estimator performs reasonably well even when u_i is distributed according to the Laplace distribution, that is, with the density $f(u) = 2^{-1}\exp(-|u|)$. Powell (1981) proposed the least absolute deviations (LAD) estimator in Tobin's model. It is defined as the value of β that minimizes $\sum_{i=1}^{n}|y_i - \max(x_i'\beta, 0)|$. He has shown the estimator to be consistent under general distributions of u_i as well as under heteroscedasticity, and derived its asymptotic distribution. The intuitive appeal for the LAD estimator in a censored regression model arises from the fact that in the case of an i.i.d. sample, the median (of which the LAD estimator is a generalization) is not affected by left censoring below the mean. A main drawback of the LAD estimator is the computational difficulty it entails. It is hoped that a reasonably efficient algorithm for computing the LAD estimator will be developed in the near future.

Gronau's Wage-Rate Model

Gronau (1973) studied the effect of children on the housewife's value of time and consequently on her wage rate, using US census data. A censored regression model is appropriate because there are many housewives who do not work and therefore for whom the wage rate is not observed.

Gronau assumed that, given the exogenously determined offered wage W^0, a housewife maximizes her utility function $U(C, X)$ subject to $X = W^0H + V$ and $C + H = T$, where C is time spent at home for child care, X represents all other goods, T is total available time, and V is other income. In Fig. 2, the budget constraint is represented by a solid line, and two possible indifference curves are drawn in broken lines.



Limited Dependent Variables, Fig. 2 A housewife's determination of hours worked

Depending on the shape of the indifference curves, there are two possible types of solutions to this maximization problem: the interior solution A or the corner solution B. In B the housewife does not work, and in A she works for H hours. To put it algebraically, the housewife does not work if

$$\left[\frac{\partial U}{\partial C}\Big/\frac{\partial U}{\partial X}\right]_{H=0} > W^0 \tag{3.1}$$

and works if the inequality above is reversed. If she works, the hours worked H are obtained by solving

$$\frac{\partial U}{\partial C}\Big/\frac{\partial U}{\partial X} = W^0, \tag{3.2}$$

and the actual wage rate W is equal to the offered wage W^0. Gronau calls the left-hand side of (3.1) the housewife's value of time or, more commonly, the reservation wage, denoted W^r. In the statistical model he estimates, Gronau is concerned only with the determination of the wage rate and not with the hours worked. (A statistical model which explains both the wage rate and the hours worked will be discussed in the next section.) Assuming that both W^0 and W^r can be written as linear combinations of independent variables plus error terms, Gronau specifies his statistical model as follows:

$$\begin{aligned} W_i^0 &= x_{i2}'\beta_2 + u_{i2} \\ W_i^r &= z_i'\alpha + v_i \\ W_i^0 &= W_i, \quad \text{if} \quad W_i^0 > W_i^r \\ &= 0, \quad \text{if} \quad W_i^0 \le W_i^r, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{3.3}$$

where (u_{i2}, v_i) are i.i.d. according to a bivariate normal distribution.

I shall write the model (3.3) in such a way that its similarity to Tobin's model (2.2) becomes more apparent. By defining $y_{i1}^* = W_i^0 - W_i^r$ and $y_{i2}^* = W_i^0$ and defining x_{i1}, β_i, and u_{i1} appropriately, I can write (3.3) equivalently as

$$\begin{aligned} y_{i1}^* &= x_{i1}'\beta_1 + u_{i1} \\ y_{i2}^* &= x_{i2}'\beta_2 + u_{i2} \\ y_{i2} &= y_{i2}^*, \quad \text{if} \quad y_{i1}^* > 0 \\ &= 0, \quad \text{if} \quad y_{i1}^* \le 0, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{3.4}$$

where (u_{i1}, u_{i2}) are i.i.d. according to a bivariate normal distribution with mean zero, variances σ_1^2 and σ_2^2 and covariance σ_{12}. One can put $\sigma_1^2 = 1$ without loss of generality.)

Like Tobin's model, model (3.4) could be used for the analysis of household expenditure on a durable good. Then, y_{i1}^* would signify an unobservable index of the intensity of the ith household's desire to buy the durable good, and y_{i2}^* would signify an unobservable index of how much the ith household wishes to spend on it. Note that Tobin's model is a special case of (3.4) obtained by assuming y_{i1}^* and y_{i2}^* are the same.

The other extreme special case of (3.4) is obtained by assuming independence between y_{i1}^* and y_{i2}^* In this special case, the computation of the ML estimators is simple: the ML estimator of β_1/σ_1 is obtained by applying the probit ML estimator of the equation for y_{i1}^*, and the ML estimator of β_2 is obtained by the least squares regression of positive y_{i2}^* on x_{i2}. Because of its computational ease, this model was often used before the advance of computer technology. However, in economic applications it is generally unrealistic to assume the independence of y_{i1}^* and y_{i2}^*.

The likelihood function of model (3.4) can be written as

$$L = \prod_0 P(y_{i1}^* \le 0) \prod_1 \int_0^\infty f(y_{i1}^*, y_{i2}) dy_{i1}^*, \quad (3.5)$$

where Π_0 denotes the product over those i for which $y_{i2} = 0$ and Π_1 over those i for which $y_{i2} > 0$. The maximization of (3.5) is relatively simple. Using (3.5), the hypothesis of the equality of y_{i1}^* and y_{i2}^*. (Tobin's hypothesis) or the hypothesis of independence between y_{i1}^* and y_{i2}^* can be tested by the likelihood ratio test or by any other asymptotically equivalent test.

Heckman's two-step estimator can be used for this model. From (3.4) we can obtain an equation analogous to (2.4) as follows:

$$E(y_{i2}|y_{i2} > 0) = x_{i2}'\beta_2 + \sigma_{12}\sigma_1^{-1}\lambda(x_{i1}'\sigma_1), \quad (3.6)$$

Where $\alpha_1 = \beta_1/\sigma_1$. In the first step one estimates α_1 by applying the probit ML estimator applied to the equation for y_{i1}^*, denoted $\widehat{\alpha}_1$, and in the second step one regresses positive y_{i2} on x_{i2} and $\lambda(x_{i1}'\widehat{\alpha}_1)$.

Heckman's estimator is consistent as long as (3.6) is valid. In particular, the consistency of Heckman's estimator does not require the joint normality of y_{i1}^* and y_{i2}^*, for (3.6) is valid as long as y_{i1}^* is normal and y_{i2}^* can be written as a sum of a linear function of y_{i1}^* and a random variable (not necessarily normal) distributed independent of y_{i1}^*.

Heckman's Labour Supply Model

The theoretical model of labour supply discussed in section "Gronau's Wage-rate Model" determines both the wage rate and the hours worked, but, as we noted earlier, Gronau's statistical model defines the distribution of only the wage rate. Heckman's (1974) statistical model which defines the joint distribution of the wage rate and the hours worked is developed as follows:

Heckman's equation for the offered wage rate (actually, its logarithm) is, like Gronau's, given by

$$W_i^0 = x_{i2}'\beta_2 + u_{i2}. \quad (4.1)$$

Heckman specifies $W^r \equiv (\partial U/\partial C)/(\partial U/\partial X)$ explicitly as a function of the hours worked H and a linear function of exogenous variables plus an error term as

$$W_i^r = \gamma H_i + z_i\alpha + v_i. \quad (4.2)$$

It is assumed that the ith individual works if

$$W_i^r(H_i = 0) \equiv z'\alpha + x_i < W_i^0 \quad (4.3)$$

and then the wage rate W_i and the hours worked H_i are determined by simultaneously solving (4.1) and (4.2) after putting $W_i^0 = w_i^r = w_i$. Therefore, we can define Heckman's statistical model as

$$W_i = x_{i2}'\beta_2 + u_{i2}$$

and

$$\begin{aligned} W_i = \gamma H_i z_i'\alpha + v_i, \quad \text{if} \quad H_i^* \equiv x_{i1}'\beta_1 + u_{i1} > 0 \\ W_i = 0 \text{ and } H_i = 0, \quad \text{if} \quad H_i^* \le 0, \end{aligned} \quad (4.4)$$

where $x_{t1}'\beta_1 = \gamma^{-1}(x_{i2}'\beta_2 + z_i'\alpha)$ and $u_{i1} = \gamma^{-1}(u_{i2} - v_i)$. Note that H_i^* may be interpreted as the desired hours of work.

The first two equations of (4.4) constitute the structural equations of a simultaneous equations model since an endogenous variable H appears in the right-hand side of the second equation. In order to make Heckman's model (4.4) comparable to Tobin's model (2.2) or Gronau's model (3.4), I shall write the reduced-form version of Heckman's model. By defining $y_i^* = H^*, y_i = H$, $y_2^* = W^0$, and $y_2 = W$, the reduced-form version of Heckman's model (assuming normality) can be defined by

$$\begin{aligned} y_{i1}^* &= x_{i1}\beta_1 + u_{i1} \\ y_{i2}^* &= x_{i2}'\beta_2 + u_{i2} \\ y_{i1} = y_{i1}^*, \quad \text{if} \quad y_{i1}^* > 0 = 0, \quad \text{if} \quad y_{i1}^* \le 0 \\ y_{i2} = y_{i2}^*, \quad \text{if} \quad y_{i1}^* > 0 = 0, \quad \text{if} \quad y_{i1}^* \le 0, \\ i = 1, 2, \ldots, n, \end{aligned} \quad (4.5)$$

where (u_{i1}, u_{i2}) are i.i.d. according to a bivariate normal distribution with mean zero, variances σ_1^2 and σ_2^2, and covariance σ_{12}.

The likelihood function of model (4.5) is given by

$$L = \prod_0 P(y_{i1}^* \leq 0) \prod_1 f(y_{i1}, y_{i2}), \tag{4.6}$$

where the definitions of the symbols are the same as in (3.5). The maximization of (4.6) is also a fairly routine problem. Heckman's two-step method can also be applied to this model in a way very similar to that discussed in section "Gronau's Wage-rate Model".

The estimation of the structural parameters of model (4.4) requires an additional procedure because of its simultaneity problem. However, as noted by Amemiya (1979), the problem of deriving the estimates of the structural parameters from the estimates of the reduced form parameters in the simultaneous equations censored regression model can be solved by means of the same principle as in the standard simultaneous equations model.

Endogenous Switching Regression Model

The endogenous switching regression model is defined as follows:

$$\begin{aligned} y_{i1}^* &= x_{i1}'\beta_1 + u_{i1} \\ y_{i2}^* &= x_{i2}'\beta_2 + u_{i2} \\ y_{i3}^* &= x_{i3}'\beta_3 + u_{i3} \\ y_i = y_{i2}^*, \quad &\text{if} \quad y_{i1}^* > 0 = y_{i3}^*, \quad \text{if} \quad y_{i1}^* \leq 0 \\ w_i = 1, \quad &\text{if} \quad y_{i1}^* > 0 = 0, \quad \text{if} \quad y_{i1}^* \leq 0, \end{aligned} \tag{5.1}$$

where (u_{i1}, u_{i2}, u_{i3}) are i.i.d. according to a trivariate normal distribution. Note that the variables with the asterisks are unobserved, and y_i, w_i, and x_is are observed.

It is called a switching regression model because the regression equation which the observed dependent variable follows switches back and forth between two equations. It is called endogenous switching because the switching is controlled by the outcome of a random variable y_{i1}^*, which may be correlated with y_{i2}^* and y_{i3}^*. The fact that w_i is observed may be characterized by the statement that the sample separation is known in this model. A variety of switching regression models arise depending on whether the switching is endogenous or exogenous and whether the sample separation is known or unknown.

The likelihood function of model (5.1) is given by

$$L = \prod_0 \int_{-\infty}^0 f_3(y_{i1}^*, y_i)\,dy_{i1}^* \prod_1 \int_0^\infty f_2(y_{i1}^*, y_i)\,dy_{i1}^*, \tag{5.2}$$

where Π_0 denotes the product over those i for which $w_i = 0$, Π_1 over those i for which $w_i = 1$, f_2 denotes the joint density of y_{i1}^* and y_{i2}^*, and f_3 the joint density of y_{i1}^* and y_{i3}^*. As in the preceding models, the ML estimation is computationally feasible, and Heckman's two-step estimation yields consistent estimates.

An interesting example of model (5.1) is given by Lee (1978), who studied the effect of union membership on the wage rate. In Lee's model, y_{i2}^* represents the logarithm of the wage rate of the ith worker in case he or she joins the union, and y_{i3}^* represents the same in case he or she does not join the union, and y_i represents the observed wage rate. Whether or not the worker joins the union is determined by the sign of the variable

$$y_{i1}^* = y_{i2}^* - y_{i3}^* + z_i'\alpha + v_i. \tag{5.3}$$

Another interesting example of model (5.1) is the so-called *disequilibrium model*, first proposed by Fair and Jaffee (1972). In their model, y_{i2}^* represents the quantity supplied, y_i is the observed quantity traded, and $y_{i1}^* = y_{i3}^* - y_{i2}^*$. Many extensions of the disequilibrium model have been proposed and the ensuing statistical problems have been discussed in the econometric literature: for a recent survey, see Quandt (1982).

Model (5.1) can be generalized to the multinomial case where the regression equation which the observed dependent variable follows switches from one to another among more than two regression equations. To get a concrete idea, I shall describe Duncan's model (1980).

He analyses the joint determination of the location of a firm and its output (for simplicity I consider a scalar dependent variable-output, but the analysis can be generalized to the case of a vector of dependent variables, as Duncan does). A firm chooses the location at which profits are maximized, and only the output at the chosen location is observed. Let s_k^i be the profit of the ith firm when it chooses the kth location, $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$, and let y_k^i be the output of the ith firm at the kth location. Then Duncan postulates

$$s_k^i = x_{k1}^{i\prime}\beta + u_k^i \tag{5.4}$$

and

$$y_k^i = x_{k2}^{i\prime}\beta + v_k^i, \tag{5.5}$$

where x_{k1}^i and x_{k2}^i are vector functions of the input–output prices and $(u_1^i, u_2^i \ldots, u_k^i, v_1^i, v_2^i, \ldots, v_k^i)$ are i.i.d. according to a $2K$-variate normal distribution. (Economic theory dictates that the same β appear in both equations.) Suppose $s_k^i > s_j^i$ for any $j \neq K$. Then a researcher observes y_k^i but does not observe y_j^i for $j \neq K$.

I shall indicate how to derive the likelihood function of Duncan's model. Let us assume $K = 3$ for simplicity. We consider a typical firm and assume that we observe the firm to choose location 1, which implies that we observe y_1 and the event characterized by the inequalities $s_1 > s_2$ and $s_1 > s_3$. (I have suppressed the superscript which would indicate this particular firm.) Therefore, the contribution of this firm to the likelihood function is

$$P(s_1 > s_2, s_1 > s_3) f\,(y_1 | s_1 > s_2, s_1 > s_3). \tag{5.6}$$

The total likelihood function is the product of these over all the firms.

Thus the likelihood function of Duncan's model consists of the probability part and the density part, like the likelihood function of the other models we have considered so far. If K is large, the maximization of the likelihood function of this model may involve a costly computation. Duncan maximizes the probability part and the density part separately and obtains two different estimates of β. Then he estimates β by using the optimal weighted average of the two estimates.

A model similar to Duncan's can be also used, for example, to analyse the joint determination of the consumer's choice of a particular brand of a durable good and the amount of expenditure on that brand; see for example Dubin and McFadden (1984).

See Also

- ▶ Censored Data Models
- ▶ Discrete Choice Models
- ▶ Labour Supply of Women
- ▶ Logit, Probit and Tobit
- ▶ Selection Bias and Self-Selection

Bibliography

Amemiya, T. 1973. Regression analysis when the dependent variable is truncated normal. *Econometrica* 41: 997–1016.

Amemiya, T. 1979. The estimation of a simultaneous-equation Tobit model. *International Economic Review* 20: 169–181.

Amemiya, T. 1985. *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Dubin, J.A., and D.L. McFadden. 1984. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52: 345–362.

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

Fair, R.C., and D.M. Jaffee. 1972. Methods of estimation for markets in disequilibrium. *Econometrica* 40: 497–514.

Gronau, R. 1973. The effects of children on the housewife's value of time. *Journal of Political Economy* 81: S168–S199.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–693.

Heckman, J.J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such model. *Annals of Economic and Social Measurement* 5: 475–492.

Lee, L.F. 1978. Unionism and wage rates: A simultaneous equations model with quantitative and limited dependent variables. *International Economic Review* 19: 415–433.

Maddala, G.S. 1983. *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.
Powell, J.L. 1981. *Least absolute deviations estimation for censored and truncated regression models*. Technical Report No.356, Institute for Mathematical Studies in the Social Sciences, Stanford University.
Quandt, R.E. 1982. Econometric disequilibrium models. *Econometric Reviews* 1: 1–63.
Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.

Limits to Growth

Wilfred Beckerman

During the 1950s and the 1960s economic growth became one of the central preoccupations of economists and economic policy makers. This was probably the result mainly of the unprecedented rates of economic growth being achieved by the advanced countries of the world, together with significant differences in the growth rates of individual countries. Hence there was considerable interest in explaining the overall acceleration of growth and the causes of the inter-country differences.

However, during the 1960s the view emerged that perhaps the high growth rate of the advanced countries was not necessarily adding commensurately to the welfare of their populations. The various reasons for this concern were first set out brilliantly, in comprehensive and persuasive terms, by E.J. Mishan (1967). Mishan enumerated various alleged disamenities of economic growth, such as pollution, congestion of travel facilities and desirable holiday resorts, and other forms of externality, as well as more spiritual effects, such as the subordination of other social values to the pursuit of commercial objectives and the consequent deterioration in society's moral standards. Mishan's highly sophisticated and articulate attack on the mindless pursuit of economic growth corresponded to growing social awareness of some of the undesirable externalities associated with economic growth – of which obvious visible pollution of various kinds, and the rise of urban violence, made an impact on many sections of the public in the more affluent countries.

At the same time concern was being expressed in some quarters about the viability of continuing high rates of growth on account of the possible resource constraints faced by the world as a whole. These concerns, together with the alleged association between economic growth and pollution, were formulated precisely in *The Limits of Growth*, a study commissioned by the 'Club of Rome' (Meadows and others 1972). This study purported to show that on any reasonable assumptions continuation of high rates of growth would mean that (i) the world would run out of resources of key materials; (ii) increasing pollution would have serious effects; and (iii) population would outrun the world's potential food supplies. This report was at first accepted by many sections of the public as constituting a scientific demonstration of the need for governments to take action to slow down growth rates.

Whilst the basic methodology used in *The Limits to Growth* was derived from that developed by Jay Forrester (1961, 1968) in that it employed a computerized 'systems dynamics' model that enabled the emphasis to be placed on the interrelationships and 'feed-backs' between different parts of a complex model, serious defects in its particular application of the Forrester techniques were immediately apparent and as soon as it appeared *The Limits to Growth* was subjected to sharp criticism by some expert commentators (see *The Economist*, 1972; Sir (now Lord) Eric Ashby 1972; Mellanby 1972; a World Bank task force 1972, H.S.D. Cole et al. on behalf of the Science Policy Research Unit of Sussex University, 1973). Its main defects included:

(1) Failure to allow for the fact that changes in the balance between demand and supply for any materials had, over the past, eventually led to changes in price which provided the stimulus, where necessary, to the discovery of new resources, to the development of substitutes, to the technological improvements in methods of exploration, extraction and refinement, to substitution in the products in which

they are embodied and so on. History is full of dire predictions that if the demand for a certain product continued to grow as before the known resources would be used up in *x* years time, and all of them have been shown by events to have been absurd. The concept of 'known resources' is a misleading one; society only 'knows' of the resources that it is worth discovering given present and prospective demands, costs, and prices.

(2) Thus the technique of inserting fixed supplies – even with some assumptions concerning eventual finite increases in these supplies – into a computer and then confronting them with indefinitely expanding demands, which must eventually overtake the supplies, bears no resemblance to the way demands and supplies have developed over the past and has no foundation in economic analysis or the particular analysis of technological innovation.

(3) Furthermore, even if the concept of 'finite resources' made sense, slower growth would not enable society to continue indefinitely: it would merely postpone the fateful day of reckoning. If resources were really 'finite' the only way the indefinite existence of society could be ensured would be to cut standards of living to infinitesimally low levels, and this did not seem to be politically feasible in democratic countries.

(4) Pollution per unit of output was being reduced and could be reduced very much more if the correct pricing policies were introduced to internalize the externalities that pollution represented. This was a problem of resource allocation at any point of time and has nothing to do with resource misallocation over time, which is what the claim that growth was excessive amounted to. Indeed, pollution tended to be worst in the poorest countries and less resources were made available to reduce pollution to optimal levels in conditions of low and slowly rising incomes.

(5) World food supplies had been rising faster than population for several decades and faster economic growth seemed to lead to slower population increases, rather than the reverse. The acute food shortages of many parts of the world reflected gross maldistribution of world food supplies. Slowing down the growth rate of the USA was not likely to increase availability of food in those parts of Africa constantly threatened by famine. If anything, insofar as it meant less aid to such countries, it would only aggravate their condition.

These and various other serious defects in *The Limits to Growth* were analysed in detail by Wilfred Beckerman (1972, 1974). As well as demonstrating fully the technical errors in the Club of Rome report, Beckerman also emphasized the elitist middle-class nature of much of the anti-growth movement. It was the middle class, he maintained, that was most conscious of losing its privileges in a rapidly growing society, and the middle classes had always been adept at presenting their own interests as a crusade for social morality fought by people of moral refinement and exquisite aesthetic sensibility, by contrast with the crass materialism of the pro-growth lobby. This appeal made some impression on idealistic youth, and on radical members of society who saw the harmful effects of growth as evidence of the evils of a profit-dominated capitalist society (in spite of the evidence that Beckerman produced concerning the even greater neglect of the environment in Soviet bloc countries). In much of this Beckerman was, of course, developing points that had been anticipated by Anthony Crosland (Crosland 1956, 1962).

The glaring errors in the Club of Rome report and the obvious partiality of the Mishan type attack on affluence, some of which were exposed, at a UN Conference on the Environment in Stockholm in 1972, by the poorer countries whose citizens were more worried about how to get a square meal next day than about the possible accumulation of sulphur dioxide in the atmosphere by the year 2050, gradually weakened the impact of the anti-growth movement. Furthermore, it had already begun to run out of steam when world economic growth was brought to a sudden halt by the first oil shock of 1973/74. And since then the rates of economic growth in the world have been very much lower than

previously. One of the consequences of this has been the emergence of mass unemployment in most of the advanced countries of the world and economic crises in many of the developing countries. Government policies to restrain demand and to reduce budget deficits in the face of increased social security payments and lower tax revenues has meant, inter alia, that expenditures on safeguarding the environment now have much lower priority than had hitherto been the case. Those sections of the population whose social consciences are most active, therefore, are now amongst those who complain most vociferously about the failure of governments to take action to accelerate economic growth. Some people are just hard to please.

See Also

- ▶ Malthus's Theory of Population
- ▶ Natural Resources
- ▶ Stagnation

Bibliography

Ashby, E. 1972a. Lecture on pollution in perspective, to the times 1000 conference. *The Spectator*, 27 May.
Ashby, E. 1972b. Pollution and the public conscience: Fifty-first Earl Gray memorial lecture. Newcastle upon Tyne: University of Newcastle.
Beckerman, W. 1972. Economists, scientists and environmental catastrophe. *Oxford Economic Papers* 24(3): 327–344.
Beckerman, W. 1974. *In Defence of economic growth*. London: Jonathan Cape. Reprinted as *Two cheers for the affluent society*. New York: St Martins, 1975.
Cole, H.S.D., et al. (eds) for the Science Policy Research Unit of Sussex University. 1973. *Thinking about the future: A critique of the limits of growth*. London: Chatto and Windus.
Crosland, A. 1956. *The future of socialism*. London: Jonathan Cape.
Crosland, A. 1962. *The conservative enemy*. London: Jonathan Cape. *The Economist*, 11 March 1972.
Forrester, J.W. 1961. *Industrial dynamics*. Cambridge, MA: MIT Press.
Forrester, J.W. 1968. *Principles of systems*. Cambridge, MA: Wright Allen Press.
Meadows, D.H., et al. 1972. *The limits to growth: A report for the Club of Rome's project on the predicament of mankind*. New York: Universe.
Mellanby, K. 1972. The phoney crisis. *Minerva* 10(3), July.
Mishan, E.J. 1967. *The costs of economic growth*. London: Staples Press.
World Bank. 1972. Report on the limits to growth. Report by a special task force of the International Bank for Reconstruction and Development (known as the World Bank), Washington, DC, September, Mimeo.

Lindahl Equilibrium

John Roberts

Abstract

Lindahl equilibrium embodies a market solution to the problem of providing public goods. Each individual faces personalized prices at which he or she may buy total amounts of the public goods. In equilibrium, these prices are such that everyone demands the same levels of the public goods and thus agrees on the amounts that should be provided. Since individuals buy the total production of public goods, the price to producers is the sum of the prices paid by individuals, and equilibrium involves the supply at these prices equalling the common demand, with costs being shared in proportion to (marginal) benefits.

Keywords

Bargaining; Efficient allocation; Externalities; Incentive compatibility; Joint production; Lindahl equilibrium; Lindahl, E. R.; Misrepresentation of preferences; Missing markets; Nash equilibrium; Optimality; Property rights; Public goods; Pure public goods; Revealed preferences; Tax incidence; Walras equilibrium

JEL Classifications

D5

Lindahl equilibrium attempts to solve the problem of determining the levels of public goods to be provided and their financing by adapting the price

system in a way that maintains its central feature of an efficient allocation being the outcome of voluntary market activities within the context of private property rights. Instead of some political choice mechanism and coercive taxation, under the Lindahl approach each individual faces personalized prices at which he or she may buy total amounts of the public goods. In equilibrium, these prices are such that everyone demands the same levels of the public goods and thus agrees on the amounts of public goods that should be provided. Since each individual buys and consumes the total production of public goods, the price to producers is the sum of the prices paid by individuals, and equilibrium involves the supply at these prices equalling the common demand. Thus, Lindahl equilibrium brings unanimity about the level of public goods provision, with costs being shared in proportion to (marginal) benefits.

The basic idea of a market solution to the problem of providing public goods is due to Erik Lindahl (1919). In its modern formulation, Lindahl equilibrium has come to play a benchmark role in the study of economies with public goods, externalities, and government expenditure which parallels that played by Walrasian competitive equilibrium in the analysis of questions where these factors are absent. For example, tax incidence can be measured relative to the Lindahl equilibrium. On the other hand, the Lindahl concept does not share the competitive equilibrium's centrality of position as a predictor of the actual outcomes of economic activity.

This latter point involves some irony, because Lindahl's original exposition of the idea treats it as having both normative and descriptive/predictive value.

Lindahl considered a legislature in which two parties represent the two homogeneous classes that constitute the electorate. (He also indicates how to extend the analysis to more classes and their representatives.) The issue is how much government activity should be carried out and how the costs of this activity should be shared between the two groups.

Lindahl identified two functions, say $f_A(s)$ and $f_B(s)$, which give, respectively, the expenditure on public activity that group A would want if it had to pay a fraction s of the corresponding costs and the level that B would want if it had to pay the complementary fraction $1 - s$. The value $x = f_A(s)$ is just the solution to the problem of maximizing the utility of after-tax income and public expenditure for group A, given that it will pay $100s\%$ of the costs, while f_B solves the corresponding problem for B. Ignoring income effects, Lindahl obtained $s = v'_A\,(f_A(s))$, where vA is A's utility for public expenditure, and, correspondingly, $1 - s = v'_B\,(f_B(s))$. Note that f_A is decreasing and f_B is increasing. Thus, assuming $f_A(0) > f_B(0)$ or $f_A(1) < f_B(1)$, so that a group bearing all the costs wants less expenditure than does the group paying nothing, there is a unique value s^* strictly between zero and one at which the two groups agree on the desired level of expenditure, that is, $x^* = fA(s^*) = f_B(s^*)$.

Much of Lindahl's analysis is in terms of bargaining between the two groups over x and s under the assumption that, at any partition of the costs, the smaller of the two proposed quantities will be implemented. (This reflects the connection to voluntary exchange, where no one is forced to transact.) He recognized that such bargaining would not automatically lead to s^*, x^*. However, he claimed that if both groups were equally adept at defending their interests, this outcome would result.

Foley (1970) provided the basic general equilibrium treatment of Lindahl's idea in the context of an Arrow–Debreu private ownership economy with both private and pure public goods (no rivalry in consumption and no possibility of exclusion) where there are zero endowments of public goods, these goods are never used as inputs, and production takes place under constant returns to scale. See Milleron (1972), Roberts (1973), and Kaneko (1977) for extensions and Roberts (1974) for a survey.

Foley's model focuses on prices for the public goods rather than cost shares. Individual demand functions for public goods, as depending on the prices of both private and public goods, are defined (exactly as for private goods) as the choices of quantities to consume that maximize utility subject to the budget constraint defined by the prices and the agent's endowment. Thus, the quantity demanded of any public good at a particular price vector differs with individual

preferences and endowments. However, the nature of pure public goods requires that all agents' consumption of any of these goods be equal. Thus, if prices are to lead different individuals all to demand the same quantities of public goods, it is clear that the prices charged to consumers must be personalized, differing across individuals to reflect differences in preferences and incomes. The price received by a producer of public goods is then the sum of the price paid by individuals, because each unit of each public good is allocated to and paid for by every individual. Meanwhile, private goods markets involve standard competitive pricing. With this, Lindahl equilibrium is a vector p of private goods prices, a vector qi of public goods prices for each consumer i, an allocation of private goods xi to each i and a vector of public goods y such that: (xi, y) is the most preferred consumption bundle for consumer i from those affordable at prices (p, qi), given i's wealth as determined by p and i's initial endowment of private goods ω_i; and also such that the net input–output vector ($\sum_i x_i - \omega_i$, y) is profit maximizing at the producer prices (p, $\sum_i q_i$). Note that both consumers and producers are following standard, competitive, price-taking behaviour just as in the Walrasian equilibrium.

Further appreciation of the connection between Lindahl and Walrasian equilibria can be gained using Arrow's insight (1970) that externalities (and the public goods problem in particular) can be viewed as a phenomenon of missing markets. Given a public goods economy with I consumers, M private goods and N public goods such as studied by Foley, consider an associated economy with I consumers, $(M + K)$ private goods, and no public goods, where $K = IN$. In this economy, each public good n in the original economy is replaced by a collection of I private goods, each of which is of interest to and consumable by only one consumer and which together are joint products in production. A net input–output vector in this economy of the form

$$(z, \tilde{y}), \quad z \in R^M, \tilde{y} = \left(y^1, \ldots, y^{IN}\right) = (y_1, y_2, \ldots, \ldots y_N, y_1, y_2, \ldots, y_N, \ldots, y_1, y_2, \ldots, y_N) \times \in R_+^{IN}$$

is producible if and only if $(z, y_1, \ldots, y_N)$ is in the production set of the original public goods economy. A Walras equilibrium in this economy is a price vector $(p, q^1, \ldots, q^{IN}) \in R_+^{MK}$ and consumption vectors $\left(x_i, y_i^1, \ldots, y_i^{IN}\right) \in R^{M+K}, i = 1, \ldots, I$, where $\left(x_i, y_i^1, \ldots, y_i^{IN}\right)$ is the most preferred bundle for i from among those costing no more than $p\omega_i$ and where $(\sum_i x_i - w_i; \sum_i y_i{}^1, \ldots, \sum_i y_i{}^{IN})$ is profit maximizing at prices $(p, q^1, \ldots, q^{IN})$. Clearly, these conditions imply $y_i^{jn}=0$, for $i \neq j$ so that no consumer receives positive amounts of another's personalized goods, and $y_i^{in} = y_j^{jn}$ for all i, j, and n, so that each individual consumes the same quantities of these personalized goods. Thus, Walras equilibria of the artificial economy exactly correspond to the Lindahl equilibria of the original economy, with a parallel correspondence between the feasible allocations in the two economies and between the Pareto optima.

This construction, which was used by Foley to prove existence of Lindahl equilibrium, illuminates the claim that the Lindahl equilibrium involves voluntary exchange in the context of maintaining private property rights. It also makes clear that Lindahl equilibria are Pareto optimal and that any optimum can be supported as an equilibrium with a reallocation of resources. (In fact, Silvestre 1984, has characterized Lindahl allocations in terms of optimality plus a condition that no agent wants to reduce his or her contribution to paying for public goods if the level of provision would be proportionately reduced.) The Lindahl equilibrium's role as a benchmark is largely attributable to its having these properties, plus the fact that the Lindahl equilibrium allocations belong to the core if blocking is defined by a group being able to produce a more preferred consumption bundle for each of its members, even if non-members contribute nothing to public goods production (Foley 1970). However, this construction also suggests some of the problems with the Lindahl equilibrium which prevent it from having great appeal as a positive prediction.

In particular, the usual complaint against a price-based solution to the public goods problem is that there would be no reason for an individual to take the Lindahl prices as given:

misrepresentation of preferences should be profitable. Of course, as long as there are only a finite number of participants in a market, the behaviour of each typically has some influence on price formation, and so the assumption of price-taking in Walrasian, private goods equilibrium is questionable too.

Progress on this incentives question requires being more specific about the mechanism used to determine the allocation as a function of the initially dispersed information about the economic environment. In this context, Hurwicz (1972) formalized the idea that there must be incentive problems even with only private goods by showing that if a mechanism always yields Pareto optima and, if participation is voluntary, so that its outcomes must be unanimously preferred to the no-trade point, then it cannot be a dominant strategy always to report one's preferences (demand) correctly. The exactly parallel result for public goods was achieved by Ledyard and Roberts (see Roberts 1976). Thus neither Walrasian nor Lindahl equilibria can be the outcome of a mechanism which is incentive compatible in this dominant-strategy sense.

Of course, the standard case in which the Walrasian equilibrium seems appealing is a 'large numbers' one where each individual's influence is small. This intuition has been formalized in a number of ways: revealing one's true demand for private goods generically is asymptotically a dominant strategy as the number of participants in the economy becomes large; only competitive allocations are in the core of large economies; Nash equilibria of various models in which individuals recognize their influence on prices converge to the competitive solution as the economy grows. However, with public goods the situation is much different: increasing the size of the economy makes price-taking less attractive. This too has been shown in various ways. Roberts (1976) showed that increasing numbers can worsen the incentives for correct revelation of preferences for public goods and that as the numbers grow, the departure of the outcome from efficiency can also increase. Muench (1972) showed that the core and Lindahl equilibria do not coincide in large economies, and Champsaur et al. (1975) demonstrated that the core of a public goods economy may actually expand when the number of consumers increases. In terms of the artificial economy, the essential intuition is that the market for each of the personalized goods is monopsonized, and the joint-product interaction constrains the bargaining power of the producer which otherwise might permit an efficient outcome to the bilateral monopoly situation. Thus, it seems that in the large numbers situations that have been the traditional concern of economics, the price-taking assumption renders the Lindahl solution of little predictive or descriptive value.

These essentially negative results are in some contrast with the results on incentives for correct revelation in iterative planning procedures for determining public goods. This literature was begun by Malinvaud (1971a, b) and Drèze and de la Vallée Poussin (1971) and is surveyed in Roberts (1986).

In this context, the notion of incentive-compatible behaviour is Nash equilibrium: each agent selects his/her responses to the central planning authority's proposals so as to maximize his/her payoff, given the strategies being used by the other agents to determine their responses. Such behaviour typically involves misrepresentation of preferences. However, various authors (Roberts 1979, Champsaur and Laroque 1982, and Truchon 1984, for example), have shown that this misrepresentation need not prevent convergence to a Pareto optimum and, in particular, to the Lindahl allocation.

However, as argued in Roberts (1986), these results are of limited interest because they rely on the implausible assumption that each agent is perfectly informed about the other's preferences. (A similar criticism can be laid against the static mechanisms for obtaining Walrasian or Lindahl allocations as Nash equilibria; Hurwicz 1979.) Moreover, once the (self-selection or truthful reporting) constraints associated with preferences being private information are recognized, it is not clear that any mechanism can achieve Lindahl allocations (see Laffont and Maskin 1979; d'Aspremont and Gerard-Varet 1979). This gives a further reason for doubting the empirical relevance of Lindahl equilibrium.

See Also

- ▶ Duality
- ▶ Incentive Compatibility
- ▶ Public Goods

Bibliography

Arrow, K.J. 1970. The organization of economic activity: Issues pertinent to the choice of market versus non-market allocation. In *Public expenditures and policy analysis*, ed. R.H. Haveman and J. Margolis. Chicago: Markham.

Champsaur, P., and G. Laroque. 1982. Strategic behavior in decentralized planning procedures. *Econometrica* 50: 325–344.

Champsaur, P., J. Roberts, and R. Rosenthal. 1975. Cores in economies with public goods. *International Economic Review* 16: 751–764.

d'Aspremont, C., and L.A. Gerard-Varet. 1979. On Bayesian incentive compatible mechanisms. In *Aggregation and revelation of preferences*, ed. J. Laffont. Amsterdam: North-Holland.

Drèze, J., and D. de la Vallée Poussin. 1971. A tâtonnement process for public goods. *Review of Economic Studies* 38: 133–150.

Foley, D. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38: 66–72.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization: A volume in honor of Jacob Marschak*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Hurwicz, L. 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *Review of Economic Studies* 46: 217–227.

Kaneko, M. 1977. The ratio equilibrium and a voting game in a public goods economy. *Journal of Economic Theory* 16: 123–136.

Laffont, J., and E. Maskin. 1979. A differential approach to expected utility maximizing mechanisms. In *Aggregation and revelation of preferences*, ed. J. Laffont. Amsterdam: North-Holland.

Lindahl, E. (1919). *Die Gerechtigkeit der Besteuerung*. Lund: Gleerup. Part I, ch. 4, 'Positive Lösung', trans. E. Henderson and reprinted as 'Just taxation – A positive solution'. In *Classics in the Theory of Public Finance*, ed. R.A. Musgrave and A.T. Peacock: Macmillan, 1958.

Malinvaud, E. 1971a. A planning approach to the public goods problem. *Swedish Journal of Economics* 11: 96–112.

Malinvaud, E. 1971b. Procedures for the determination of a program of collective consumption. *European Economic Review* 2: 187–217.

Milleron, J. 1972. Theory of value with public goods: A survey article. *Journal of Economic Theory* 5: 419–477.

Muench, T. 1972. The core and the Lindahl equilibrium of an economy with a public good: An example. *Journal of Economic Theory* 4: 241–255.

Roberts, J. 1973. Existence of Lindahl equilibrium with a measure space of consumers. *Journal of Economic Theory* 6: 355–381.

Roberts, J. 1974. The Lindahl solution for economies with public goods. *Journal of Public Economics* 3: 23–42.

Roberts, J. 1976. The incentives for correct revelation of preferences and the number of consumers. *Journal of Public Economics* 6: 359–374.

Roberts, J. 1979. Incentives in planning procedures for the provision of public goods. *Review of Economic Studies* 46: 283–292.

Roberts, J. 1986. Incentives, information and iterative planning. In *Information, incentives, and economic mechanisms*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Silvestre, J. 1984. Voluntariness and efficiency in the provision of public goods. *Journal of Public Economics* 24: 249–256.

Truchon, M. 1984. Nonmyopic strategic behavior in the MDP planning procedure. *Econometrica* 52: 1179–1189.

Lindahl on Public Finance

Peter Böhm

Erik Lindahl (1891–1960) was one of the pioneers of modern theory of public finance. His single most important contribution is, perhaps, his treatment of the problem of 'just' taxation. He showed that, by systematic application of the so-called benefit principle, a significant part of this problem can be subjected to scientific analysis. Furthermore, his work in public finance paved the way for integrating public goods into general equilibrium models of the market economy.

Lindahl's international reputation in this field stems mainly from two writings on public finance both of which were originally published in German. They date as far back as 1919 and 1928, but did not attract general attention until the 1950s, then probably as a consequence of the rising interest in public finance that accompanied the growth

of the public sector in most countries after World War II. In 1958 central parts of these writings were made available in English translation (Musgrave and Peacock 1958).

At the time when Lindahl's two major works in public finance were written, taxation and public expenditure were to a large extent treated as separate theoretical problems. The leading theory of taxation was based on the ability-to-pay principle, whereas public expenditure was typically regarded as determined on unified theory. The cornerstone of this theory was the value premise that expenditure as well as taxes should be determined by the size of the benefits which individuals derived from public expenditure.

Elements of this benefit approach to the two central issues in public finance had been enunciated by U. Mazzola, E. Sax and a handful of other, mainly Italian, economists already in the 1880s. Around the turn of the century, Knut Wicksell (1896) emerged as a prominent advocate of this approach. These economists argued that taxes should be regarded as voluntary payments for public services in correspondence with individual preferences for these services. This approach was consistent with the subjective theory of value, developed during the latter part of the 19th century, and implied that the provision of public services was put on the same footing as the satisfaction of demand for private commodities by the market economy. But 'the final statement' of this voluntary-exchange doctrine, and the implicit equilibrium concept with taxes treated as prices for publicly provided services, was given by Lindahl (Musgrave 1959).

Lindahl's theory of taxation and public expenditure, first presented in his doctoral dissertation (1919) and elaborated in (1928), was cast in the tradition of his mentor, Knut Wicksell. Thus Lindahl assumed that issues concerning the income distribution were determined on a purely political basis and should therefore be separated from a scientific analysis of the principles of public expenditure and taxation. In his analysis of these principles, Lindahl set out from an assumption of a given, just income distribution.

Government expenditure, in Lindahl's model, referred only to so-called public or collective goods. He argued that individual taxes, which contributed to the financing of these goods, could be viewed in the same way as implicit prices for joint products (cotton and cotton seed served as one of Lindahl's examples). In equilibrium the implicit prices for joint products would reflect differences in relative demand for each of the products. Analogously, different individuals or groups of homogeneous individuals who jointly consume public goods would have different demand prices or willingness to pay for a marginal unit of government expenditure. Hence, with individual tax shares interpreted as prices for marginal units of government expenditure, each individual would prefer government activity to be extended up to the point where his marginal willingness to pay for the services produced equalled his tax share. An equilibrium could be said to exist if tax shares differed among individuals according to their different marginal willingness to pay and if the level of government expenditure was so chosen that all tax shares added up to the marginal unit of government expenditure. This concept of a public-sector equilibrium, elaborated versions of which were later called a Lindahl equilibrium, conforms to the concept of market equilibria for private goods under perfect competition, specifically for jointly produced private goods.

Lindahl's application of the benefit principle to taxation was intended to determine the optimum level of government expenditure and of total taxes as well as the tax rates for individual taxpayers. As tax rates in equilibrium reflected the marginal benefits received, Lindahl argued that the resulting taxation should be regarded as just. Furthermore, he argued that the benefit approach incorporated the ability-to-pay approach as well, since in equilibrium, marginal benefits received could be said to reflect marginal ability to pay. Thus the larger the marginal benefits received and, hence, the larger the marginal ability to pay, the higher the tax share.

This latter point is one on which Lindahl has been criticized. The basis for this particular

criticism is that the ability-to-pay principle is intended to correct for an initially unjust income distribution; this problem was, as noted, assumed away in the Lindahl model. Lindahl has also been criticized for interpreting his model as determining unambiguously an optimum level of taxes and expenditure, in spite of the fact that the introduction of a public sector in an initial situation of a just income distribution could be expected to give (widely) different total net benefits to different individuals. Thus there remains the question of whether these net benefits should be redistributed among individuals to obtain a just real income distribution in the final position. If so, a redistribution could be expected to affect the equilibrium levels of taxes and expenditure. In other words, the assumption that a just income distribution is initially given does not take care of all the distribution problems and allows more than one equilibrium position.

Lindahl was cautious when advancing his model as a description of how budgetary issues were, or could be, settled in the real world. Specifically, he observed that political power might differ among groups of individuals, hence obstructing the attainment of an equilibrium position. Still, he held the view that in a long-run perspective his model performed fairly well in explaining voter behaviour, interaction among interest groups and actual government decision-making with respect to public expenditure and tax structure. However, in real-world economies, where a large number of different kinds of taxes are used and a large number of public goods are produced by government, voter evaluation of marginal increases in government expenditure will hardly harmonize with voter behaviour as assumed by Lindahl, not even as a long-run approximation. Today, the empirical relevance of Lindahl's model is made even more problematical by the often considerable amount of taxes raised for income redistribution purposes. The benefit principle in general and Lindahl's model in particular may carry more weight as a description of the budgetary process in a future of extensive two-way communication between government and its constituents. This, however, requires a practicable solution to the problem of making people reveal their preferences for public goods, a problem generally believed to follow from the so-called free-rider incentive.

As long as such a solution is lacking, demand revelation will remain a major stumbling-block for practical use of the benefit approach. This caveat, although observed by Wicksell (1896), seemed to have escaped Lindahl.

Lindahl's other contributions to public finance concern various topics in taxation. In particular, his discussion of the tax base for income taxation (1933) has gained international recognition (Break 1954). But also in this case, it has been the theoretical rigour of his analysis more than its practical applicability that has won general acclaim.

See Also

- ▶ Public Goods
- ▶ Revelation of Preferences

Bibliography

Break, G.F. 1954. Capital maintenance and the concept of income. *Journal of Political Economy* 62, February, 48–62.

Lindahl, E. 1919. *Die Gerichtigkeit der Besteuerung. Eine Analyse der Steuerprinzipien auf Grundlage der Grenznutzenteorie*. Lund: Gleerup The central part of Lindhal's theory of public finance is published in English translation in Musgrave and Peacock (1958).

Lindahl, E. 1928. Einige strittige Fragen der Steuerteorie. In *Die Wirtschaftsteorie der Gegenwart*, ed. H. Mayer, Vol. IV. Vienna: J. Springer Published in English translation in Musgrave and Peacock (1958).

Lindahl, E. 1933. The concept of income. In *In economic essays in Gustav Cassel*. London: G. Allen & Unwin.

Lindahl, E. 1959. Om skatteprinciper och skattepolitik. *Ekonomi Politik Samhälle*, 151–71. Trans: Tax principles and tax policy. *International Economic Papers* 10: 7–23, 1960.

Musgrave, R.A. 1959. *Theory of public finance*. New York: McGraw Hill.

Musgrave, R.A., and A.T. Peacock, ed. 1958. *Classics in the theory of public finance*. London/New York: Macmillan.

Wicksell, K. 1896. *Finanztheoretische Untersuchungen*. Jena: G. Fischer.

Lindahl, Erik Robert (1891–1960)

Otto Steiger

Abstract

Erik Lindahl's writings between 1919 and 1959 covered four major areas. In public finance his pioneering contribution is today known as the 'Wicksell–Lindahl paradigm of just taxation'. In dynamic analysis he was first to develop the methods of intertemporal equilibrium and temporary equilibrium. In macroeconomics Lindahl anticipated many of the insights of Keynes's *General Theory,* and in his discussion of the concepts of income and capital he laid the foundations of the theory of national accounting. His contributions in the four fields have been acknowledged internationally step by step only since the 1950s – in the third area since the 1970s.

Keywords

Accelerationist hypothesis; Aggregate demand and supply; Benefit rule; Budget balance; Capital theory; Central banking; Davidson, D.; Debreu, G.; Domar, E. D.; European Central Bank; European Monetary Union; Eurosystem; *Ex ante* and *ex post*; Expectations; Financial equilibrium; Friedman, M.; Frisch, R. A. K.; General equilibrium; Hansen, B.; Hayek, F. A.; Hicks, J. R.; Imperfect foresight; Income; International Economic Association; Intertemporal equilibrium; Kaldor, N.; Keynes, J. M.; Leijonhufvud, A.; Lindahl, E. R.; Lundberg, E. F.; Monetary equilibrium; Monetarist(s); Myrdal, G.; Musgrave, R. A.; National accounting; Normal rate of interest; Neumann, J. von; Ohlin, B. G.; Peacock, A. T.; Phelps, E. S.; Paradox of saving; Perfect foresight; Public debts; Public expenditure; Public finance; Public works; Quantity theory of money; Riksbank; Samuelson, P.A.; Sequence analysis; Stockholm School; Stock and flow; Swedish Unemployment Committee; Taxation; Temporary equilibrium; Time; Total capital; Unemployment equilibrium; Wicksell–Lindahl paradigm of just taxation; Wicksell, J. G. K.

JEL Classifications

B31

Lindahl was born on 21 November 1891 in Stockholm and died on 6 January 1960 in Uppsala, Sweden. He is now reckoned one of the great economists who were at work between the two world wars, and earned his reputation above all as a leading member within a group of Swedish economists during the 1930s consisting, besides himself, of Gunnar Myrdal, Bertil Ohlin, Dag Hammarskjöld, Alf Johansson, Erik Lundberg and Ingvar Svennilsson –a body which Ohlin (1937) had baptised the 'Stockholm School'.

The son of a prison governor, Lindahl grew up in Jönköping, the capital of a province in southern Sweden. After passing the studentexamen at a Stockholm Secondary School in spring 1910, he enrolled the following autumn as a student at the University of Lund, where economics soon became the favourite subject in his studies of humanities and law, which he passed with the degrees of the fil*osofie kandidatexamen* (BA) in 1912 and the *juris kandidatexamen* (LLB) in 1914. Although Knut Wicksell was professor of economics and fiscal law in Lund at that time (1901–16), Lindahl did not have any personal contact with him during this period. However, Emil Sommarin, the successor to Wicksell's chair (1916–39) and at the time of Lindahl's student years *docent* (reader) in economics and a great admirer of Wicksell, succeeded in encouraging Lindahl to study the former's works to such an extent that the latter became in effect the first pupil of Wicksell. As Lindahl's dissertation of 1919, *Die Gerechtigkeit der Besteuerung,* was largely based on Wicksell's theory of public finance (1896), Sommarin let Wicksell read and comment on it, and, at the public defence of the thesis at Lund University on 13 December 1919, Wicksell officiated as the official 'challenger' appointed by the faculty of law (Lindahl 1951, pp. 26–7).

With his doctoral thesis Lindahl had earned the title *docent* in public finance at Lund University (1920) and later also in economics and fiscal law at Uppsala University (1924), but not yet the position of a professor. In 1926 he became responsible for the planning of the voluminous investigations on *Wages, Cost of Living and National Income in Sweden 1860–1930* (see Lindahl 1937a; Benny Carlson 1982, pp. 11–20) carried on in the following decade at the Institute for Social Sciences in Stockholm University and financed by the Rockefeller Foundation. In his attempts to obtain a chair in economics Lindahl failed twice: in 1924 he lost the competition for a professorship at the University of Copenhagen to Bertil Ohlin, later his colleague in the Stockholm School, and in 1930 he was ranked as number two only for a chair in political economy and sociology at Gothenburg University, this time defeated by Gustaf Åkerman, like Lindahl an early pupil of Wicksell.

Only two years later, however, in 1932, Lindahl obtained the chair in political economy at the Gothenburg School of Business Economics without application, and from this time onwards Swedish universities competed to call him to their departments of economics. In 1939, the year of publication of his most famous work, *Studies in the Theory of Money and Capital,* he succeeded Sommarin at Lund University, and in 1942 he became professor at the University of Uppsala, where he retired in 1958. Internationally, Lindahl's outstanding position as economist was honoured by his election as President of the International Economic Association in 1956.

Lindahl's growing reputation from the early 1930s onwards also led to numerous calls for economic expertise by Sweden's governments and official institutions. When Sweden left the gold standard in 1931 he became an adviser to Riksbanken, the Swedish central bank. When as a result of the Great Depression the final report of the Swedish Unemployment Committee had to be given a theoretical foundation of its proposal for public works as remedy against unemployment (see Hammarskjöld 1935, p. ix and ch. 1; Otto Steiger 1971, p. 40; and Bent Hansen 1981, pp. 266–7) and when, therefore, the character of Sweden's budget system had to be superseded in 1937 by a system deliberately designed to operate in a countercyclical manner (see Lindahl 1935; cf. 1939a, app.), his expertise was sought by the Minister of Finance. Lindahl also became an economic adviser to the League of Nations (1936–9) and on two occasions to the United Nations (1949–50 and 1952–4).

Lindahl's work can be said to cover four major areas: (*a*) public finance; (*b*) methods of dynamic analysis; (*c*) monetary and macroeconomic theory; and (*d*) concepts of income and capital. Although Lindahl did not neglect empirical research, especially in public finance and national accounting, his contributions concentrated mainly on pure economic theory (see the detailed bibliography by Gertrud Lindahl and Olof Wallmén 1960).

Public Finance

Lindahl started his scientific career with a treatise on 'just taxation', his doctoral dissertation of 1919, which built on Wicksell (1896) and which, together with two re-examinations in 1928 and 1959, made it a pioneering contribution to the economic theory of the public household, today known as the ' Wicksell–Lindahl paradigm of just taxation' (Heinz Grossekettler 2006, p. 557; for more detail see Peter Bohm 1987, and John Roberts 1987). It can be characterized as a culmination of the neoclassical reformulation of the classical version of the benefit approach to the simultaneous determination of public revenue and expenditure – a reformulation which applied a new interpretation of the benefit rule as a condition of equilibrium instead of as a standard of justice as in the classical version.

Lindahl formulated this condition in a partial equilibrium framework, where 'financial equilibrium', that is, the equilibrium of *public* finance, is determined by equalization of the ratio of prices paid by each taxpayer for public and for private goods to his marginal benefits derived from public and from private goods, the equilibrating financial process brought about by the political mechanism in a parliamentary democracy (cf. Roberts 1987). Lindahl was convinced that his model could explain voting behaviour and the influence of

pressure groups on decisions of the government concerning public expenditures and taxes (Bohm 1987, p. 201). However, this 'voluntary exchange approach' (Richard A. Musgrave 1959, pp. 73–8) for a long time failed to meet with much understanding. The importance of Lindahl's path-breaking contribution was first acknowledged in the 1950s via the works on the pure theory of public expenditure of Paul A. Samuelson (1954) and Musgrave (1959) as well as by the English translation of important parts of his dissertation of 1919 in 1958 in the volume *Classics in Public Finance,* edited by Musgrave and Alan T. Peacock. In the 1970s and the 1980s, however, Lindahl's model came under attack. It was criticized for relying on the ' implausible assumption that each agent is perfectly informed about the other's preferences' (Roberts 1987, p. 200), and also for lacking empirical relevance in face of today's, unlike in Lindahl's time, 'considerable amount of taxes raised for income distribution purposes' (Bohm 1987, p. 201).

Methods of Dynamic Analysis

Lindahl's contributions to dynamic method were formulated as part of the theoretical core of his macroeconomic ideas, culminating in 1939 in his *Studies in the Theory of Money and Capital.* As has been shown by Björn Hansson (1982; cf. 1987, 1991, pp. 168–202; and Jan Petersson 1987), Lindahl's dynamic theory was developed by mutual influence within the Stockholm School, with himself and Myrdal as the key figures and mainly independent not only from influences from other contemporary economists but also – contrary to William P. Yohe (1959) – from Wicksell.

Already in his first macroeconomic treatise, the first edition of *Penningpolitikens mål* ([The aims of monetary policy], 1924, ch. 3), Lindahl stressed the *time* factor as a problem for economic analysis and used the notion of 'subjective calculations of the future' and also the term *ex post* (p. 33). A first coherent dynamic method was formulated in his treatise on capital theory (1929a; cf. 1939a, pt. III), where Lindahl developed the famous notion of *intertemporal equilibrium,* that is, the analysis of the sequential character of an economy by a sequence of periods with equilibrium in each period as a consequence of the assumption of perfect foresight. This approach has been praised by Gérard Debreu (1959, p. 35) as being 'the first mathematical study of an economy whose activity extends over a finite number of elementary time-intervals'. However, as has been pointed out later (Murray Milgate 1982, pp. 133–5), Friedrich A. Hayek had been moving on similar lines one year earlier. But this does not disturb the claim of Lindahl's originality, because a comparison of the 1929 and 1930 editions of his *Penningpolitikens medel* [The means of monetary policy] clearly shows that Lindahl became aware of Hayek's approach first after having worked out his own concept – Hayek's (1928) paper is referred to only in the second (1930, p. 11), not in the first edition (1929c, p. 10).

As has been shown by Hansson (1982, ch. 4, 59–67; 1987, pp. 504–5), Lindahl's formulation of intertemporal equilibrium, however, does not really represent a sequential process, since all prices and quantities are determined simultaneously at the beginning of the process for all periods. Lindahl became aware of this weakness when, under the influence of Myrdal's explicit introduction of expectations in equilibrium theory (1927, ch. 1), in the last section of his treatise on capital theory (1929a, pp. 80–1; cf. 1939a, pt. III, pp. 348–50) he substituted imperfect for perfect foresight. In *Penningpolitikens medel* (1930, pp. 18–24, 31–2; cf. 1939a, pt. II, pp. 158–9) Lindahl abandoned therefore, for the case of imperfect foresight, the method of intertemporal equilibrium for the notion of *temporary equilibrium*, that is, the analysis of the sequence of an economy as a series of very short periods of temporary equilibria with changes allowed only at the transition points of the periods. This notion looks closely akin to John R. Hicks's dynamic analysis in *Value and Capital* (1939, ch. 9) which in fact had been influenced decisively by Lindahl via personal contacts in 1934 and 1935, as later acknowledged by Hicks (cf. 1973, p. 8; 1985, pp. 66, 69; 1991, pp. 372–6; and Claes-Henric Siven 2002, pp. 142–5). More important in a historical perspective, however, is the striking fact

that general equilibrium theorists, from the late 1960s onwards, began to give up their mathematically more elaborated intertemporal equilibrium models of Arrow–Debreu type and to develop different notions of temporary equilibrium for very much the same reason as Lindahl in 1929 and 1930: recognition of the fact that intertemporal equilibrium does not reflect the sequential character of an economy in an essential way and the impossibility of handling imperfect foresight, that is, problems involving uncertainty and money.

However, under the influence of the criticism of his approach by Lundberg in 1930 and Myrdal in 1932 and 1933, Lindahl realized that even with the notion of temporary equilibrium there was no real causation between the periods when he applied this dynamic method to the analysis of the saving–investment mechanism during a Wicksellian cumulative process, since the equilibrium approach in the construction of temporary equilibrium cannot handle unforeseen events *during* a period. Therefore, he abandoned this notion and formulated instead the method of *sequence analysis.* This was done in the first part, Section 1, of his *Studies* (1939b, pp. 21–69), but a fully developed sequence analysis had already been presented in two unpublished papers of 1934a (published in Steiger 1971, pp. 204–11) and 1935 (cf. Hansson 1982, ch. 9). Furthermore, in Section 2 of the first part of his book Lindahl was the first economist who, in an extensive algebraic discussion of the relations between fundamental economic concepts (1939b, pp. 74–136), made the methodologically important distinction between 'micro-economic' and 'macroeconomic terms' (p. 74) by which he tried to base the relations between macro values on some kind of microeconomic behaviour (pp. 111, 125; cf. Svennilsson 1938, ch. 1; Siven 1991, pp. 155–6; and Jens Christopher Andvig 1991, p. 414). As shown by Hal R. Varian (1987, p. 461), this innovation has been wrongly attributed to Ragnar Frisch who, in an article of 1933 (pp. 172–3), had used the related terms 'micro-dynamic' and 'macro-dynamic analysis' in which he, however, 'was uninterested in the problems of microeconomic roots' of macroeconomics (Andvig 1991, p. 415).

Incorporating the method of *ex ante* and *ex post* (cf. Steiger 1987a) developed by Myrdal (1932, 1933) in his disequilbrium analysis and adopted by Ohlin (1934, ch. 1), where the former had criticized Lindahl's method of temporary equilibrium, and taking care of the sequence analysis of consecutive periods formulated by Hammarskjöld (1933a, 1933b, chs 1–5) and Svennilsson (1938, ch. 1), Lindahl's dynamic method in 1939b consisted of two parts: (a) a single-period analysis where *ex ante* plans determine *ex post* results; and (b) a continuation analysis where these *ex post* events lead to revised *ex ante* plans of a subsequent period. While Lindahl allowed for disequilibrium as long as he analysed a single period only, his analysis for several periods demanded equilibrium within each period. Because of this assumption Lindahl's sequence analysis – although it can be regarded as the first dynamic method with a meaningful sequential character, that is, not relying on the mutual interdependence of all events – did not imply the solution to the dynamic problem of establishing a convincing explanation of the causal connection between successive periods. In the end, while acknowledging Myrdal's plea for disequilibrium analysis, Lindahl hesitated to rely on the 'cumbersome *ex ante* and *ex post* terminology' (1939b, p. 68; cf. 1939c, pp. 264–5) because of its 'analytical complexity' (Siven 2006b, p. 694; cf. 1985, p. 590; Hansen 1981, p. 274).

It was left to Lundberg's sequence analysis of 1937 (ch. 9) to overcome this limitation by allowing for disequilibrium within the different periods with the help of the assumption of constant expectation functions (cf. Hansson 1982, ch. 10).

Lindahl accepted Lundberg's method in his Studies (1939b, pp. 57–9), but was not keen on the time-related model sequences based on difference equations which were incorporated in the latter's construction. On the contrary, this dynamic method was rejected by Lindahl because of its mechanical character resulting from the assumption that expectations need not enter explicitly. However, it was exactly this approach which came to dominate dynamic theory until the late 1960s when general equilibrium theorists

reintroduced the notion of temporary equilibrium and developed dynamic models which look very similar to Lindahl's sequence analysis (for example, Frank H. Hahn 1980).

Monetary and Macroeconomic Theory

While Lindahl's contributions to dynamic analysis were formulated independently of Wicksell, his work on monetary and macroeconomic theory was clearly derived from the latter (1898, 1906). This influence can be traced back as far as Lindahl's first treatise on monetary matters, *Penningpolitikens mal* (1924, 1929b), where he systematized and extended the concepts used in the Swedish controversy between Wicksell and David Davidson before and after the First World War on the aim of monetary policy being to preserve the real value of contracts, that is, Wicksell's desideratum of a constant price level versus Davidson's proposal of price level variations in inverse proportion to changes in productivity (cf. Hammarskjöld 1944; Carl G. Uhr 1960, pp. 270–305; Klas Fregert 1993; and Siven 2002, 124–9). While Lindahl's analysis is worked out along Wicksellian lines, he ends up with Davidson's and not Wicksell's solution by showing that the latter's proposition of a *'normal' rate of interest,* that is, the particular level of the money rate which is equal to the 'natural' rate, determined by the marginal productivity of capital, does not hold for a constant price level in the face of productivity variations.

A more systematic treatment of Wicksell's concept of the normal rate was given in *Penningpolitikens medel* (1930, pp. 121–30; cf. 1939a, pt. II, pp. 245–57), where Lindahl was the first to show that this notion implies three different conditions for equilibrium: '(1) it corresponds to the *natural* or . . . the *real rate of interest*; (2) it establishes *equilibrium between the demand for and supply of saving* [that is, investment and savings]; and (3) it is *neutral in* relation to *the price level* – whereas a rate of interest above or below "normal" will influence the price level in a downward or upward direction' (1939a, pt. II, p. 246; cf. 1930, p. 122). It was this formulation which inspired Myrdal's famous reconstruction of Wicksell's normal rate (1932, 1933, 1939) in which the three conditions were characterized as *monetary* equilibrium (cf. Steiger 1987c, p. 507; Siven 2006a, pp. 11–12; 2006b, pp. 672–4).

However, in the central part of *Penningpolitikens medel*, the analysis of the relation between the rate of interest and the price level, Lindahl (1930, pp. 131–4; cf. 1939a, pt. II, pp. 257–60; and 1939c, pp. 260–8) did not employ the notion of the normal rate, because his explicit consideration of expectations showed him the impossibility of a unique equilibrium rate irrespective of the rate of change of the level of prices - a reasoning very similar to John Maynard Keynes's emphasis in the *General Theory* (1936, pp. 242–4) that there are different normal rates for different levels of employment. Instead, Lindahl explained changes in the general price level with the help of another concept introduced by Wicksell (1906, p. 159): the approach of *aggregate demand and supply.* In Lindahl's formulation of this approach changes in the price level were determined by changes in the relation between the total demand for and the total supply of consumption goods, the total demand for consumption goods being defined as 'the portion of the total nominal income which is not saved', $E\ (1 - s)$ where E denotes total nominal income and s the ratio of saving to income, and the total supply defined as PQ, where P denotes the price level and Q the quantity of consumer goods of a certain period (1930, pp. 12–13; cf. 1939a, pt. I, pp. 142–3). In general, he never analysed imbalances in macroeconomic variables 'caused by "wrong" relative prices' (Siven 2002, p. 141). Using Wicksell's suggestion of a perfect credit system, this approach left no room for the quantity of money either, and it was indeed, in Lindahl's analysis of the issue of money by the central bank, directed against the quantity theory of money, although he did not deprive it of all significance for the theory of money (cf. 1929d, p. 18; 1955, pp. 32–4). In fact, as has been first emphasized by Hansen (1979, p. 123) and later confirmed by Axel Leijonhufvud (1991, p. 464) and Lars Werin (1991, p. 178) as well as Mauro Boianovsky and Hans-Michael Trautwein (2006, pp. 881–2, 888–95), Lindahl in his later writings

(cf. 1957, pp. 13–15, 19–21) was 'a true monetarist' and the first one to formulate the 'accelerationist' hypothesis in inflation theory offered a decade later by Edmund S. Phelps (1967) and Milton Friedman (1968). This hypothesis lies also at the heart of Lindahl's critical reformulation of *The General Theory,* in which he criticized Keynes's method of comparative statics in the equilibration of savings and investment by changes in income and employment, because it presupposes 'correct anticipations' (Lindahl 1953, p. 27; cf. already 1934a, pp. 209–10; 1939c, pp. 264–5). Instead, Keynes should have relied on the dynamic analysis of changes of monetary and real variables where, like in his own analysis, expectations are allowed to adapt to the changes, but where expectational errors may nevertheless have effects that alter the equilibrium rates of interest and (un)employment (cf. Boianovsky and Trautwein 2006, p. 897). It is most interesting to note that Keynes (1934), in a correspondence with Lindahl (1934b) on the latter's paper of 1934a, rejected Lindahl's method because its 'dealing with time leads to undue complications and will be very difficult either to apply or to generalise about'.

Although Lindahl did not attempt to explicitly explain changes in output and employment, his aggregate analysis resulted in achievements which paved the way for Myrdal's (1932, 1933, 1939) and Ohlin's (1933, 1934, chs 1–3) monetary approaches, and which are still important for modern macroeconomics: (*a*) the use of the savings ratio *s* in the expression E(1 – s) which related saving to expected income and which can be regarded as an alternative formulation of Keynes's (1936) propensity to consume, led to a definite distinction between saving and investment, with Lindahl (1929c, pp. 11–12, written in 1927–28; cf. Hansen 1981, p. 261) being the first economist to see the independence of the latter from the former variable; *(b)* their distinction allowed him to divide aggregate income into saving and consumption demand and aggregate output into investment and consumer goods; (*c*) the 'paradox of savings' could be solved according to which a reduction in savings results in increased savings; (*d*) the assumption of unused resources was introduced (1930, pp. 42–51; cf. 1939a, 176–9 and 185–6), and unemployment was explained by deflation caused by a fall of aggregate demand where even the possibility of a stable unemployment equilibrium was visualized (1929c, pp. 43–4; 1930, p. 44); however, deleted in 1939a, pt. II; cf. Hansen 1981, pp. 261–3). Compared with Keynes's principle of effective demand determining the equilibrium level of (un) employment there are, however, certain limitations in Lindahl's aggregate demand/ supply approach: (*a*) unemployment equilibrium was considered only as an 'exceptional case', with – like in the other Stockholm School analyses on the relation between unemployment and wages (esp. Alf Johansson 1934, ch. 5) – 'rigid money wages as a necessary condition for' and no 'complete macro model of unemployment' (Hansen 1981, pp. 268–9; cf. Siven 2002, p. 141); (*b*) the rate of interest was treated in its orthodox role as equilibrator of savings and investment in the long run; (*c*) saving and investment were not equilibrated by changes in aggregate income but by variations in its *distribution* (cf. the discussion initiated by Karl-Gustav Landgren 1960, ch. 6:3; and followed up by Steiger 1971, pp. 173–9; 1978, pp. 424–5; 1991, 129–30; Hansen 1981, pp. 261–3; Don Patinkin 1982, pp. 44–6; and Johan Myrman 1991, pp. 272–6). In the analysis of this adjustment process, however, Lindahl was able to anticipate the whole neo-Keynesian or Kaldor–Pasinetti theory of distributive shares (cf. Guglielmo Chiodi and Kumaraswamy Velupillai 1983; Velupillai 1988).

On the other side, the equilibrating role of changes in total income with respect to saving and investment is implicit in Lindahl's sequence analysis of 1934 (1934a, pp. 208–11), where he showed how a difference between investment and saving *ex ante* leads to their *ex post* equality, and it was clearly visualized in his discussion of loan-financed public works as a means against unemployment (1932, pp. 136–7; 1935, pp. 1–5; cf. 1939a, app., pp. 356–67). As has been pointed out by Hansen (1955, p. 41), Lindahl in *Penningpolitikens medel* (1930, pp. 63–8; however, deleted in 1939a, pt. II) was the first economist to consider the possibility of a systematic use

of variations in the relation between public expenditures and public incomes, that is, the budget balance, as a means to stabilize economic fluctuations because, as he recognized, a surplus in the balance can be defined as equivalent to state saving and a deficit as state investment (1930, p. 65). With this analysis he paved the way for Ohlin's (1934, ch. 5) and Myrdal's (1934, pts III–IV) more detailed analysis of loan- financed public works as remedies against unemployment. Unlike Keynesian economists, however, Lindahl in this discussion did not neglect the effects of such a fiscal policy for the national debt which he analysed in detail in 1944 (cf. 1946). There he formulated rules governing state borrowing which are very similar to those developed at the same time by Evsey D. Domar (1944), that is, that the problem of public debts is first and foremost a problem of the growth of national income.

Another innovation in Lindahl's monetary and macroeconomic theory was his discussion of how to organize the central banking system in a monetary union of independent nations (cf. 1930, pp. 170–9; however, deleted in 1939a, pt. II). As recently recognized by Gunnar Heinsohn and Steiger (2003, p. 13; cf. Steiger 2007, pp. 43–5), Lindahl was the first economist to develop the model of a decentralized, two-stage central banking system for a common currency consisting of a main central bank and the national central banks, where the latter would receive the banknotes in the same way from the former like the domestic commercial banks of their national central bank. With this model he hoped to open the possibility for the main central bank to equilibrate differences in real rates of interest due to different rates of inflation between the union's members by allowing for differences in nominal rates of interest. With this proposal, Lindahl anticipated the central Achilles' heel of the Eurosystem, the central banking system of the European Monetary Union (EMU), where its Council can decide only on a unique nominal rate of interest (cf. Dieter Spethmann and Steiger 2005, pp. 55–8). Furthermore, in spite of its name, the European Central Bank is not designed for issuing money and, therefore, not the central monetary authority of the EMU that could push through such a differentiation of credit. In his model, Lindahl (1930, p. 171) also demonstrated the necessity of a central fiscal authority in a monetary union to support the main central bank – another problem that has not been solved in the EMU (cf. Heinsohn and Steiger 2003, pp. 13, 39).

Concepts of Income and Capital

While Lindahl's contributions to macroeconomic theory have been discussed extensively in the literature on the Stockholm School, his work on the macroeconomic concepts of income and capital have not received much attention (cf. Yohe 1962).

The discussion of the notions of capital and income in Lindahl's theoretical framework stemmed from two different roots: (a) the approach to capital theory conceiving capital goods as stored services of land and labour, originally formulated by Eugen von Böhm-Bawerk (1889) and developed by Wicksell (1893, 1901) and Gustaf Åkerman (1923–4); and (*b*) the approach to the concept of income regarding income as a flow of benefits from the stock of capital and introduced into economic theory by Irving Fisher (1906). Both approaches had in common that *time* was included in a decisive manner and in concentrating on this element, Lindahl made the notions of income and capital essentially correlative.

In his piece on capital theory where he had introduced the method of intertemporal equilibrium (1929a, 1939a, pt. III), Lindahl avoided the theoretical difficulties of working with the concept of total capital in a world of heterogeneous capital goods by developing a completely disaggregated stationary equilibrium system – a 'Walrasian model with capital à la Wicksell' and with 'a striking similarity' to John von Neumann's model of equilibrium growth of 1937 (Hansen 1970, pp. 199, 207–8). Although Lindahl did not make use of the concept of total capital in his equilibrium system, it can be shown that its total capital value can be determined on the basis of the term 'income' employed in his model – an insight which Lindahl formulated unequivocally in

Penningpolitikens medel (1930, pp. 13–15; cf. 1939a, pt. II, pp. 143–6) and in more detail in his contributions on the concept of income (1933, 1937b, pp. 76–111).

Starting from Irving Fisher's basic premise that income consists of the services obtained from capital goods during a certain period, whereas capital is a stock existing at a given point of time, Lindahl (1933, pp. 400–1) looked upon *income as interest* accruing on the value of capital goods, and considered capital value as future income discounted. With this concept of income, implying that income is equal to the sum of consumption *and* saving, he solved the inconsistencies in Fisher's analysis of capital and income where saving, a flow term expressing the increase in capital value, had been excluded from income and incorporated into capital, a stock term. Consequently, Lindahl's discussion also made clear that changes in capital value, contrary to what had been the premises of Böhm-Bawerk, Wicksell and Åkerman, are not determined by changes in the use of capital but in the use of income, that is, the part which is not consumed: saving.

However, as Lindahl realized, this thesis holds true only under the assumption of perfect foresight. Following Myrdal's analysis of expectations of 1927, he showed that as soon as uncertainty about future events is introduced changes in anticipation of owners of capital assets lead to changes in capital value by *gains* and *losses* which cannot be regarded as positive or negative income, because like the stock of capital they refer not to a certain period but to a point of time (1929a, p. 75; 1939a, pt. III, p. 341; cf. Myrdal 1927, p. 44, and the further discussion in Lindahl, 1939b, pp. 101–10). This insight led Lindahl to abandon the most practical concept of income – income as earnings – because it included gains and losses. Although Lindahl's concept of income – like his abstract classification of capital goods (Hansen 1970, p. 200) – has been criticized for not being capable of empirical application and measurement, his contributions – together with Myrdal's approach of 1927 – have been acknowledged as 'the fundamental theoretical work concerning the notion of income' (Nicholas Kaldor 1955, p. 162). This work should also become fundamental to Lindahl's research on national accounting (1937b, 1939b, pp. 74–136; 1954) which made him 'the father of Social Accounting *theory'* (Hicks 1973, p. 8; cf. Carlson 1982, 33–6).

During his lifetime – and until the late 1980s – Lindahl never earned a reputation comparable to that of his colleagues in the early Stockholm School, Gunnar Myrdal and Bertil Ohlin, and it has been argued by one of his younger colleagues (Lundberg 1982, p. 275) that his contributions to economics were not distinguished by 'ingenious ideas' like those of Myrdal and Ohlin. As has been shown in this survey of Lindahl's work, however, the numerous original ideas in each of the four fields covered by his writings argue for quite different judgement. This holds especially true for Lindahl's monetary and macroeconomic theory, as has been demonstrated since the late 1980s by Boianovsky and Trautwein (2006), Leijonhufvud (1991), Siven (1991, 2002, 2006a, 2006b), Steiger (1987a, 1987b, 1987c, 2006a, 2006b, 2007), and Velupillai (1988).

The author wishes to thank Claes-Henric Siven (Stockholms Universitet) and Hans-Michael Trautwein (Universität Oldenburg) for valuable suggestions.

See Also

- ▶ Ex Ante and Ex Post
- ▶ Lindahl Equilibrium
- ▶ Myrdal, Gunnar (1898–1987)
- ▶ Ohlin, Bertil Gotthard (1899–1979)
- ▶ Stockholm School

Selected Works

1919. *Die Gerechtigkeit der Besteuerung. Eine Analyse der Steuerprinzipien auf der Grundlage der Grenznutzentheorie.* Lund: Gleerupska Universitetsbokhandeln and H. Ohlsson. Ch. 4 trans. as 'Just taxation – a positive solution', in Musgrave and Peacock (1958).

1924. *Penningpolitikens mål och medel. Del I* [The aims and means of monetary policy. Part

I]. Lund: C.W.K. Gleerup; Malmö: Försäkringsaktiebolaget. 1st edn of Lindahl (1929b).
1928. Einige strittige Fragen der Steuertheorie. In *Die Wirtschaftstheorie der Gegenwart*. Vol. 4, ed. H. Mayer. Vienna: J. Springer; abridged trans. as 'Some controversial questions in the theory of taxation', in Musgrave and Peacock (1958).
1929a. Prisbildningsproblements uppläggning från kapitalteoretisk synpunkt [The formulation of the theory of prices from the viewpoint of capital theory]. *Ekonomisk Tidskrift* 31: 31–81. Revised version trans. as Pt. III of Lindahl (1939a).
1929b. *Penningpolitikens mål* [The aims of monetary policy]. zund: C.W.K. Gleerup; Malmö: Försäkringsaktiebolaget. 2nd edn of Lindahl (1924).
1929c. *Penningpolitikens medel* [The means of monetary policy]. Lund: C.W.K. Gleerup; Malmö: Försäkringsaktiebolaget. 1st edn of Lindahl (1930).
1929d. *Om forhallåndet mellan penningmängd och prisnivån* [On the relation between the quantity of money and the price level]. Uppsala: Lundequistska and Almqvist & Wiksell.
1930. *Penningpolitikens medel* [The means of monetary policy]. Lund: Gleerup; Malmö: Försökringsaktiebolaget. 2nd edn of Lindahl (1929c); revised version trans. as Pt. II of Lindahl (1939a).
1932. Offentliga arbeten i depressionstider [Public works in times of depression]. *Nationalekonomiska Föreningens Förhandlingar* [Proceedings of the Swedish Economic Association], Meeting of 25 November, 127–137, 163–164.
1933. The concept of income. In *Economic essays in Honour of Gustav Cassel.* 20 October 1933. London: Allen & Unwin.
1934a. A note on the dynamic pricing problem. Mimeo, Gothenburg, 13 October. Quoted from the corrected version published in Steiger (1971).
1934b. Letter to John Maynard Keynes. Gothenburg, 7 November. In Steiger (1971).
1935. Arbetslöshet och finanspolitik [Unemployment and fiscal policy]. *Ekonomisk Tidskrift* 37: 1–36. Revised version trans. as App. to Lindahl (1939a).
1937a. (with E. Dahlgren and K. Kock). *National income of Sweden 1861–1930. In two parts.* Vol. III of *Wages, Cost of Living and National Income in Sweden 1860–1930,* ed. the Staff of the Institute for Social Sciences, University of Stockholm. London/Stockholm: P.S. King & Son/P.A. Norstedt & Söner.
1937b. 'National income', the concept and methods of estimation. In Lindahl (1937a), Pt. I.
1939a. *Studies in the theory of money and capital.* London: Allen & Unwin.
1939b. The dynamic approach to economic theory. Pt. I of Lindahl (1939a).
1939c. Additional note (1939) (to Lindahl, 1939a, Pt. II). In Lindahl (1939a).
1944. Teorien för den offentliga skuldsättningen [The theory of the public debt]. In *Studier i ekonomi och historia tillägnade Eli* F *Heckscher* [Studies in economics and history in honour of Eli F. Heckscher]. Uppsala: Almqvist & Wiksell. Abridged version trans. as Lindahl (1946).
1946. The problem of the growing national debt. *Skandinaviska Banken Quarterly Review* 27(2): 43–48. Abridged version of Lindahl (1944).
1951. Till hundraårsminnet av. Knut Wicksells födelse [On the centenary of Knut Wicksell's birth]. *Ekonomisk Tidskrift* 53: 197–243. Quoted from the abridged version trans. as 'Wicksell's Life and Work', in K. Wicksell, *Selected papers on economic theory*, ed. E. Lindahl. London: Allen & Unwin, 1958.
1953. Om Keynes' ekonomiska system. *Ekonomisk Tidskrift* 55: 186–243. Quoted from and trans. as 'On Keynes' economic system I–II', *Economic Record* 30 (1954): 19–32 and 159–171.
1954. Nationalbokföringens grundbegrepp. *Ekonomisk Tidskrift* 56: 87–138. Trans. as 'The basic concepts of national accounting', *International Economic Papers* 7 (1957): 71–100.
1955. Penningteoretiska utgångspunkter [Starting points in monetary theory]. In *Om Riksbankens*

sedelutgivningsrätt och därmed sammanhängende frågor [On the Riksbanken's right to issue banknotes and related questions]. Stockholm: Statens offentligar utredningar 1955: 43, 32–57.

1957. *Spelet om penningvärdet.* Stockholm: Kooperativa Förbundet. Quoted from and trans. as 'Das Spiel mit dem Geldwert', *Weltwirtschaftliches Archiv* 87 (1961): 7–53.

1959. Om skatteprinciper och skattepolitik. In *Ekonomi, politik, samhälle. En bok tillägnad Bertil Ohlin* [Economics, politics, society. A publication in honour of Bertil Ohlin], Stockholm: Folk & Samhalle and Esselte. Trans. as 'Tax principles and tax policy', *International Economic Papers* 10 (1960): 7–23.

Bibliography

Åkerman, J.G. 1923–4. *Realkapital und Kapitalzins.* 2 vols. Stockholm: Centraltryckeriet.

Andvig, J.C. 1991. Ragnar Frisch and the Stockholm School. In Jonung (1991).

Bohm, P. 1987. Lindahl on public finance. In *The New Palgrave. A dictionary of economics.* Vol. 3, ed. J. Eatwell, M. Milgate and P. Newman. London: Macmillan.

Boianovsky, M., and H.-M. Trautwein. 2006. Price expectations, capital accumulation and employment. Lindahl's macroeconomics from the 1920s to the 1950s. *Cambridge Journal of Economics* 30: 881–900.

Carlson, B. 1982. Bagge, Lindahl och nationalinkomsten. Om 'National Income of Sweden 1861–1930' [Bagge, Lindahl and the national income. On 'National Income of Sweden 1861–1930']. *Meddelande från Ekonomisk–historiska institutionen, Lunds Universitet,* No. 27.

Chiodi, G., and K. Velupillai. 1983. A note on Lindahl's theory of distribution. *Kyklos* 36: 103–111.

Debreu, G. 1959. *Theory of value. An axiomatic analysis of economic equilibrium*. New York: Wiley.

Domar, E.D. 1944. The 'burden of the debt' and the national income. *American Economic Review* 34: 798–827.

Fisher, I. 1906. *The nature of capital and income.* New York: Macmillan. 2nd ed., 1912.

Fregert, K. 1993. Erik Lindahl's norm for monetary policy. In *Swedish economic thought. Explorations and advances*, ed. L. Jonung. London: Routledge.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Frisch, R. 1933. Propagation problems and impulse problems in economic dynamics. In *Economic essays in honour of Gustav Cassel. October 20th, 1933*. London: Allen & Unwin.

Grossekettler, H. 2006. Wicksell, Knut: Finanztheoretische Untersuchungen (1896). In *Lexikon der ökonomischen Werke. 650 wegweisende Schriften von der Antike bis ins 20. Jahrhundert*, ed. D. Herz and V. Weinberger. Düsseldorf: Verlag Wirtschaft und Finanzen.

Hahn, F.H. 1980. Unemployment from a theoretical viewpoint. *Economica NS* 47: 285–298.

Hammarskjöld, D. 1933a. Utkast till en algebraisk metod för dynamisk prisanalys [Outline of an algebraic method for the dynamic analysis of prices]. *Ekonomisk Tidskrift* 34 (5–6) (1932, printed 1933): 157–176.

Hammarskjöld, D. 1933b. *Konjunkturspridningen. En teoretisk och historisk undersökning* [The propagation of business cycles. A theoretical and historical investigation]. Stockholm: Statens offentliga utredningar 1933: 29.

Hammarskjöld, D. 1935. *Åtgärder mot arbetslöshet. Arbetslöshetsutredningens betänkande, del II* [Measures against unemployment. Report of the Swedish Unemployment Committee, Part II]. Stockholm: Statens offentliga utredningar 1935: 6.

Hammarskjöld, D. 1944. Den svenska diskussionen om penningpolitikens mål. In *Studier i ekonomi och historia tillägnade Eli F. Heckscher* [Studies in economics and history in honour of Eli F. Heckscher]. Uppsala: Almqvist & Wiksell. Trans. as 'The Swedish discussion on the aims of monetary policy'. *International Economic Papers* 5 (1955), 145–154.

Hansen, B. 1955. *Finanspolitikens ekonomiska teori.* Stockholm: Statens offentliga utredningar, 1955: 25. Quoted from and trans. as *The economic theory of fiscal policy.* London: Allen & Unwin, 1958.

Hansen, B. 1970. *A survey of general equilibrium systems*. New York: McGraw-Hill.

Hansen, B. 1979. Review of Erik Lundberg (ed.): Inflation theory and anti-inflation policy (1977). *Scandinavian Journal of Economics* 81: 119–125.

Hansen, B. 1981. Unemployment, Keynes, and the Stockholm School. *History of Political Economy* 13: 256–277.

Hansson, B. 1982. *The Stockholm School and the development of dynamic method*. London: Croom Helm.

Hansson, B. 1987. Stockholm School. In *The New Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 4. London: Macmillan.

Hansson, B. 1991. The Stockholm School and the development of dynamic method. In *The history of Swedish economic thought*, ed. B. Sandelin. London: Routledge.

Hayek, F.A. 1928. Das intertemporale Gleichgewichtssystem der Preise und die Bewegungen des 'Geldwertes'. *Weltwirtschaftliches Archiv* 28, 33–76. Trans. as 'Intertemporal price equilibrium and movements in the value of money', in F.A. von Hayek, *Money, capital and fluctuations. Early essays,* ed. R. Cloughry. London: Routledge & Kegan Paul, 1984.

Heinsohn, G. and O. Steiger 2003. The European Central Bank and the Eurosystem. An analysis of the missing

central monetary institution in the European Monetary Union. ZEI Working Paper No. B03–09. Center for European Integration Studies (ZEI), Universität Bonn. Repr. in *The Euro, the eurosystem, and the European economic and monetary union*, ed. D. Ehrig and O. Steiger. Hamburg: LIT-Verlag, 2007.

Hicks, J.R. 1939. *Value and capital. An inquiry into some fundamental principles of economic theory.* Oxford: Clarendon Press. 2nd ed, 1946.

Hicks, J.R. 1973. Recollections and documents. *Economica NS* 40: 2–11.

Hicks, Sir J. 1985. *Methods of dynamic economics*. Oxford: Oxford University Press.

Hicks, Sir J. 1991. The Swedish influence on *value and capital.* In Jonung (1991).

Johansson, A. 1934. *Löneutvecklingen och arbetslösheten* [The development of wages and unemployment]. Stockholm: Statens offentliga utredningar 1934: 2.

Jonung, L., ed. 1991. *The Stockholm School of economics revisited*. Cambridge: Cambridge University Press.

Kaldor, N. 1955. *An expenditure tax*. London: Allen & Unwin. Quoted from the reprint of 54–78, 'The concept of income in economic theory', in *Readings in the concept and measurement of income,* ed. R.H. Parker and G.C. Harcourt. Cambridge: Cambridge University Press, 1969.

Keynes, J.M. 1934. Letter to Erik Lindahl. Bloomsbury, 8 December. Quoted from the reproduction in Steiger (1971).

Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.

Landgren, K.-G. 1960. *Den 'nya ekonomien' i Sverige. J.M. Keynes, B. Ohlin och utvecklingen 1927–39* [The 'new economics' in Sweden. J.M. Keynes, B. Ohlin and the development 1927–39]. Stockholm and Uppsala: Almqvist & Wiksell.

Leijonhufvud, A. 1991. Roundtable discussion. In Jonung (1991).

Lindahl, G., and O. Wallmén. 1960. Erik Lindahl. Bibliograf! 1919–1960 [Erik Lindahl. Bibliography 1919–1960]. *Ekonomisk Tidskrift* 62: 59–74.

Lundberg, E. 1930. Om begreppet ekonomisk jämvikt [On the notion of economic equilibrium]. *Ekonomisk Tidskrift* 32: 133–160.

Lundberg, E. 1937. *Studies in the theory of economic expansion*. Stockholm: Norstedt & Söner.

Lundberg, E. 1982. Lindahl, Erik. In *Svenskt biografiskt lexikon* [The Swedish Dictionary of Bibliography] 23. Stockholm: Norstedts Tryckeri.

Milgate, M. 1982. *Capital and employment. A study of Keynes's economics*. London: Academic Press.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

Musgrave, R.A. and A.T. Peacock. eds. 1958. *Classics in the theory of public finance.* London: Macmillan. 2nd ed, 1967.

Myrdal, G. 1927. *Prisbildningsproblemet och föränderligheten* [The problem of price formation and changeability]. Uppsala and Stockholm: 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 doctrine]. *Ekonomisk Tidskrift* 33 (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 rev. version of Myrdal (1932); 2nd rev. version trans. as Myrdal (1939).

Myrdal, G. 1934. *Finanspolitikens ekonomiska verkningar* [The economic effects of fiscal policy]. Stockholm: Statens offentliga utredingar 1934: 1.

Myrdal, G. 1939. *Monetary equilibrium.* London: Hodge. Revised version of Myrdal (1933).

Myrman, J. 1991. The monetary economics of the Stockholm School. In Jonung (1991).

Ohlin, B. 1933. Till frågan om penningteoriens uppläggning [On the question of the method and structure in monetary theory]. *Ekonomisk Tidskrift* 35, 45–81. Quoted from and translated as 'On the formulation of monetary theory', *History of Political Economy* 10 (1978), 353–388.

Ohlin, B. 1934. *Penningpolitik, offentliga arbeten, subventioner och tullar som medel mot arbetslöshet. Bidrag till expansionens teori* [Monetary policy, public works, subsidies and tariffs as remedies for unemployment. A contribution to the theory of expansion]. Stockholm: Statens offentliga utredningar 1934: 12.

Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment. I–II. *Economic Journal* 47: 53–69, 221–240.

Patinkin, D. 1982. *Anticipations of the general theory? And other essays on Keynes*. Chicago: University of Chicago Press.

Petersson, J. 1987. *Erik Lindahl och Stockholsskolans dynamiska metod* [Erik Lindahl and the dynamic method of the Stockholm School]. Lund: Universitetsförlaget Dialogus.

Phelps, E.S. 1967. Phillips curves, expectations of inflation and optimal unemployment over time. *Economica NS* 34: 254–281.

Roberts, J. 1987. Lindahl equilibrium. In *The New Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London/New York: Macmillan/Stockton Press.

Samuelson, P.A. 1954. The pure theory of public expenditures. *Review of Economics and Statistics* 36: 387–389.

Siven, C.-H. 1985. The end of the Stockholm School. *Scandinavian Journal of Economics* 87: 577–593.

Siven, C.-H. 1991. Expectation and plan. The microeconomics of the Stockholm School. In Jonung (1991).

Siven, C.-H. 2002. Analytical foundations of Erik Lindahl's monetary analysis, 1924–1930. *History of Political Economy* 34: 111–153.

Siven, C.-H. 2006a. Stockholmsskolan och Keynes [The Stockholm School and Keynes]. *Okonomisk Forum* 60 (6): 10–17.

Siven, C.-H. 2006b. Monetary equilibrium. *History of Political Economy* 38: 665–709.
Spethmann, D. and O. Steiger 2005. The four Achilles' heels of the Eurosystem. Missing central monetary institution, different real rates of interest, non-marketable securities, and missing lender of last resort. *International Journal of Political Economy* 34(2), (2004, printed Summer 2005) 46–68.
Steiger, O. 1971. *Studien zur Entstehung der Neuen Wirtschaftslehre in Schweden. Eine Anti-Kritik*. Berlin: Duncker & Humblot.
Steiger, O. 1978. Prelude to the theory of a monetary economy. Origins and significance of Ohlin's 1933 approach. *History of Political Economy* 10: 420–446.
Steiger, O. 1987a. *Ex ante* and *ex post*. In *The New Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London/New York: Macmillan/Stockton Press.
Steiger, O. 1987b. Lindahl, Erik Robert (1891–1960). In *The New Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London/ New York: Macmillan/Stockton Press.
Steiger, O. 1987c. Monetary equilibrium. In *The New Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London/ New York: Macmillan/Stockton Press.
Steiger, O. 1991. Comment on Eskil Wadensjö: 'The Committee of Unemployment and the Stockholm School'. In Jonung (1991).
Steiger, O. 2006a. Lindahl, Erik Robert: *Penningpolitikens medel* [The means of monetary policy] (1930). In *Lexikon der ökonomischen Werke. 650 wegweisende Schriften von der Antike bis ins 20. Jahrhundert*, ed. D. Herz and V. Weinberger. Düsseldorf: Verlag Wirtschaft und Finanzen.
Steiger, O. 2006b. Lindahl, Erik Robert: *Studies in the theory of money and capital* (1939). In *Lexikon der ökonomischen Werke. 650 wegweisende Schriften von der Antike bis ins 20. Jahrhundert*, ed. D. Herz and V. Weinberger. Düsseldorf: Verlag Wirtschaft und Finanzen.
Steiger, O. 2007. Erik Lindahl och Eurosystemet [Erik Lindahl and the eurosystem]. *Ekonomisk Debatt* 35 (2): 42–45.
Svennilsson, I. 1938. *Ekonomisk planering. Teoretiska studier* [Economic planning. Theoretical studies]. Uppsala: Almqvist & Wiksell.
Uhr, C.G. 1960. *Economic Doctrines of Knut Wicksell*. Berkeley/Los Angeles: University of California Press.
Varian, H.R. 1987. Microeconomics. In *The New Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London: Macmillan.
Velupillai, K. 1988. Somer Swedish stepping stones to modern macroeconomics. *Eastern Economic Journal* 14: 87–98.
von Böhm-Bawerk, E.1889. *Kapital und Kapitalzins. Zweite Abteilung: Positive Theorie des Kapitales.* Jena: G. Fischer; 4th ed, 1921. Trans. as *Capital and interest,* vols. 2 and 3. South Holland: Libertarian Press, 1959.
von Neumann, J. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Browerschen Fixpunktsatzes. In *Ergebnisse eines mathematischen Kolloquiums* 8, ed. K. Menger. Trans. as 'A model of general equilibrium', *Review of Economic Studies* 13 (1945–6), 1–9.
Werin, L. 1991. There were *two* Stockholm Schools. In Jonung (1991).
Wicksell, K. 1893. *Über Wert, Kapital und Rente.* Jena: G. Fischer. Trans. as *Value, capital and rent,* London: Allen & Unwin, 1954.
Wicksell, K. 1896. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens.* Jena: G. Fischer, iv–vi, 76–87, 101–59. Trans. as 'A new principle of just taxation', in Musgrave and Peacock (1958).
Wicksell, K. 1898. *Geldzins und Güterpreise. Eine Studie über die den Tauschwert des Geldes bestimmenden Ursachen. Jena: G. Fischer. Trans. as Interest and prices. A study of the causes regulating the value of money.* London: Macmillan, 1936.
Wicksell, K. 1901. *Föreläsningar i nationalekonomi.* Vol. 1, Stockholm: Fritzes; Lund: Berlingska. Trans. of the 3rd Swedish edn (1928) as *Lectures on political economy. volume 1: General theory,* London: Routledge & Sons, 1934.
Wicksell, K. 1906. *Föreläsningar i nationalekonomi*. Vol. 2: *Om penningar och kredit*. Stockholm: Fritzes; Lund: Berlingska. Quoted from the trans. of the 3rd Swedish edn (1929), *Lectures on political economy. volume 2: Money,* London: Routledge & Sons, 1935.
Yohe, W.P. 1959. *The Wicksellian tradition in Swedish macroeconomic theory.* Ann Arbor: University Microfilms.
Yohe, W.P. 1962. A note on some lesser known works of Erik Lindahl. *Canadian Journal of Economics and Political Science* 28: 274–280.

Linear Models

Edwin Burmeister

JEL Classifications
C6

Von Neumann's linear economic model was first published in German in 1938 and translated into English in 1945. Since then there have been numerous economic and mathematical refinements, most

of which are summarized in Burmeister and Dobell (1970) and/or Morishima (1969). The original von Neumann formulation did not admit either primary factors or final consumption. However, in the generalized von Neumann model described below, one primary factor, labour, is allowed, as well as a vector of final consumption goods. A further generalization allowing a vector of different primary factors is possible. Accordingly, linear models of Leontief–Sraffa (1960) type become a special case.

Assume there exist m alternative production activities for producing n different commodities, where $m \lessgtr n$. Activity j operating at a unit intensity level requires a labour input a_{0j} and a vector of commodity inputs $(a_{1j}, a_{2j}, \ldots, a_{nj})$ to produce a vector of commodity outputs $(b_{1j}, b_{2j}, \ldots, b_{nj})$. Production takes one time period, so inputs at time t result in outputs at time $t+1$. Constant returns to scale is implied by linearity, and inputs λa_{0j} and $(\lambda a_{1j}, \lambda a_{2j}, \ldots, \lambda a_{nj})$ yield outputs $(\lambda b_{1j}, \lambda b_{2j}, \ldots, \lambda b_{nj})$ for all $\lambda \geq 0$.

The vector of labour requirements for the m alternative production activities is written as $A_0 = (a_{01}, a_{02}, \ldots, a_{0m})$. The input matrix is

$$\begin{bmatrix} A_0 \\ \hline A \end{bmatrix} = \begin{bmatrix} a_{01} & - & a_{0m} \\ a_{11} & & a_{1m} \\ \vdots & & \vdots \\ a_{n1}, & \ldots & , a_{nm} \end{bmatrix}$$

the output matrix is

$$B = \begin{bmatrix} b_{11} & \ldots & b_{1m} \\ \vdots & & \vdots \\ b_{11} & \ldots & b_{nm} \end{bmatrix}$$

and the intensity levels at which each of the m production activities is operated is given by the column vector

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

Although in general some of the a_{0j}'s, a_{ij}'s and b_{ij}'s may be zero, here we assume that they are all positive; we thereby avoid some technical difficulties and gain expositional simplicity.

Assume that labour grows at the rate $g \geq 0$,

$$L(t+1) = (1+g)L(t), \tag{1}$$

and is fully employed with

$$L(t) = A_0(t)x(t). \tag{2}$$

For all t production must satisfy the resource constraint

$$Ax \leq Bx - C \tag{3}$$

where C denotes the column vector of commodities consumed

$$C = \begin{bmatrix} C_1 \\ \vdots \\ C_m \end{bmatrix}$$

The economy is capable of producing positive final consumption and balanced growth at the rate g provided the inequalities

$$(1+g)Ax \leq Bx - C, \quad \begin{matrix} C \geq 0, & C \neq 0 \\ x \geq 0, & x \neq 0 \end{matrix} \tag{4}$$

are satisfied.

Prices for a unit of labour services and the n commodities are given by $p_0 = w$ and the row vector $p=(p_1, p_2, \ldots, p_n)$, respectively, where

$$\sum_{i=0}^{n} p_i = 1$$

is one convenient normalization. The steady-state (or balanced growth) rate of interest (or profit rate) is denoted by r.

The economy can achieve a steady-state equilibrium at a given value of $r \geq 0$ if the von Neumann price system has a solution satisfying the inequalities

$$wA_0 + (1+r)pA \geq pB, \quad \begin{matrix} w \geq 0, & w \neq 0 \\ p \geq 0, & p \neq 0. \end{matrix} \tag{5}$$

L

The quantity system (4) is dual to the price system (5) when $r = g$.

The von Neumann solution to (5) involves three economically essential inequalities:

(i) If the cost of operating an activity exceeds the revenue from that activity, then that activity is not used in a steady-state equilibrium solution, that is that activity is operated at a zero intensity level. Formally,

$$x_j = 0 \text{ if } wa_{0j} + (1+\mathrm{r})\sum_{i=1}^{n} p_i a_{ij} > \sum_{i=1}^{n} p_i b_{ij}, \quad j = 1, \ldots, m. \tag{6}$$

(ii) If a commodity has a positive price, then its supply and demand are equal:

$$(1+g)\sum_{i=1}^{m} a_{ij}x_j = \sum_{j=1}^{m} b_{ij}x_j - C_i \quad \text{if } p_i > 0, i = 1, \ldots, n. \tag{7}$$

(iii) The price of a commodity is zero if it is in excess supply:

$$p_i = 0 \text{ if } (1+g)\sum_{i=1}^{m} a_{ij}x_j < \sum_{j=1}^{m} b_{ij}x_j - C_i, i = 1, \ldots, n. \tag{8}$$

The above generalized von Neumann model is not a general equilibrium model because there is one missing equation. Some behavioural equation involving the rate of interest and consumption is required to form a general equilibrium model. Nevertheless, this incomplete specification can be used to confirm and generalize many well known results.

For example, in steady-state equilibrium Eqs. (5), (6), (7) and (8) imply that

$$(1+g)Ax + pC = pBx = \text{value of output} = wA_0 + p(1+r)Ax. \tag{9}$$

Using (2), the per capita value of commodity inputs or capital is given by

$$v = \frac{pAx}{A_{0x}}; \tag{10}$$

similarly the per capita value of consumption is

$$pc = \frac{pC}{A_0 x}. \tag{11}$$

Substituting (10) and (11) into (9) and rearranging yields

$$pc = w + (r-g)v. \tag{12}$$

The Golden Rule result that the value of per capita consumption is equal to the wage rate at the Golden Rule point where $r = g$ is an immediate consequence of (12). Other such results are easily derived; thus allowing for the possibility of joint production does not invalidate many economic results.

Two familiar classes of models are special cases of this generalized von Neumann model. First, Leontief–Sraffa models result simply by setting $m = n$ and $B = I$, which implies that the technology is free of joint production. Then if $p > 0$ from (7)

$$C = [I - (1+g)A]x \tag{13}$$

where now the column vector x is interpreted as the output vector for commodities 1,..., n. Provided (4) has a solution, (13) may be solved for

$$x = [I - (1+g)A]^{-1}C, \tag{14}$$

and premultiplying (14) by the vector A_0 gives the consumption possibility frontier

$$A_0 x = L = [I - (1+g)A]^{-1}C. \tag{15}$$

Similarly, steady-state equilibrium prices for the Leontief–Sraffa model are given by

$$p = wA_0[I - (1+r)A]^{-1}. \tag{16}$$

Second, most neo-Austrian models of the type studied by Hicks (1973a) are a special case of this generalized von Neumann model. The latter fact is most easily demonstrated by considering the simple numerical example due to Burmeister (1974).

A neo-Austrian process is a time sequence of input–output vectors

$$\{(a_t, b_t)\}_{t=0}^{T} \tag{17}$$

where a_t is the input of a commodity and b_t is the output (of the same commodity) in period t. Consider a process

$$\{(a_t, b_t)\}_{t=0}^{2} = \{(a_0, 0), (a_1, b_1), (a_2, 1)\}. \tag{18}$$

The neo-Austrian model (18) is *equivalent* to a von Neumann specification with

$$\left[-\frac{A_0}{A}-\right] = \left[\begin{array}{cccc} a_1 & a_1 & a_2 & \\ - & - & - & - \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & \end{array}\right] \tag{19}$$

and

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & b_1 & 1 \end{bmatrix}; \tag{20}$$

see Burmeister (1974, pp. 441–4).

We see, therefore, that the generalized von Neumann model is extraordinarily useful for unifying several apparently different ways of describing the production technology. However, when we do not restrict our attention to steady-state equilibria, the dynamic evolution of the model becomes extremely complex. The inequalities (4) and (5) must be satisfied for each t, as well as some additional equation to determine the interest rate r. Known results on the dynamics of models with heterogeneous capital goods – see, for examples the discussion and references cited in Chapters 5 and 6 of Burmeister (1980) – warn us that the task of completely characterizing the dynamic properties of von Neumann models will not be easy. The fact that the von Neumann formulation admits joint production makes the task even harder.

See Also

- ▶ Hawkins–Simon Conditions
- ▶ Input–Output Analysis
- ▶ Linear Programming
- ▶ Marxian Value Analysis
- ▶ Non-substitution Theorems
- ▶ Perron–Frobenius Theorem
- ▶ Sraffian Economics

Bibliography

Burmeister, E. 1974. Synthesizing the neo-Austrian and alternative approaches to capital theory: A survey. *Journal of Economic Literature* 12 (2): 413–456.

Burmeister, E. 1980. *Capital theory and dynamics*. New York: Cambridge University Press.

Burmeister, E., and A.R. Dobell. 1970. *Mathematical theories of economic growth*. New York: Macmillan.

Burmeister, E., and K. Kuga. 1970. The factor-price frontier, duality and joint production. *Review of Economic Studies* 37 (109): 11–19.

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

Hicks, J.R. 1973b. The Austrian theory of capital and its rebirth in modern economics. In *Carl Menger and the Austrian school of economics*, ed. J.R. Hicks and W. Weber. New York/London: Oxford University Press.

Morishima, M. 1969. *Theory of economic growth*. Oxford: Clarendon Press.

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

von Neumann, J. (1945–6). Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. In *Ergebnisse eines mathematischen Kolloquiums*, vol. 8, ed. K. Menger. Leipzig: Verlag. Trans. by G. Morgenstern as 'A model of general economic equilibrium'. *Review of Economic Studies* 13: 1–9. Reprinted in *Readings in mathematical economics,* ed. P. Newman, vol. II. Baltimore: Johns Hopkins University Press, 1968, 221–229.

von Weizsacker, C.C. 1971. *Steady state capital theory*. New York: Springer.

Linear Programming

George B. Dantzig

Abstract

An article by George Dantzig, the 'father of linear programming'. The problem of minimizing or maximizing a function of several variables subject to constraints when all the

L

functions are linear is called a 'linear program'. Linear programs can be used to approximate the broad class of convex functions commonly encountered in economic planning. Thousands of linear programs are efficiently solved with the simplex method, an algorithm. Solving a model with alternative activities requires software not only for solving on computers large systems of equations but also for selecting the best combination from an astronomical number of possible combinations of activities.

Keywords

Bimatrix games; Convex program; Dantzig, G.; Decomposition principle; Dantzig, G. B.; Kantorovich, L. V.; Koopmans, T. C.; KuhnTucker conditions; Lagrange multipliers; Leontief input–output model; Linear programming; Mathematical programs; Mini-max theorem; Mixed strategies; Simplex method for solving linear programs; von Neumann, J.

JEL Classifications

C6

A list of applications of linear programming, since it was first proposed in 1947 by G. Dantzig, could fill a small volume. Both J. von Neumann and L. Kantorovich made important contributions prior to 1947. Its first use by G. Dantzig and M. Wood was for logistical planning and deployment of military forces. A. Charnes and W. Cooper in the early 1950s pioneered its use in the petroleum industry. S. Vajda and E.M.L. Beale were early pioneers in the field in Great Britain. In socialist countries, it is used to determine the plan for optimal growth of the economy. Thousands of linear programs are efficiently solved each day all over the world using the simplex method, an algorithm, also first proposed in 1947. Many problems which once could only be solved on high-speed mainframe computers can now be solved on personal computers.

The problem of minimizing or maximizing a function f_0 of several variables $X = (X_1, X_2, \ldots, X_n)$ subject to constraints $f_i(X) \leq 0$, $i = 1, \ldots, n$ is called a *mathematical program.* When all the functions f are linear, it is called a *linear program;* otherwise a *non-linear program.* If all *f* are convex functions, it is called a *convex program*. At first glance, linear inequality systems appear to be a very restricted class. However, as pointed out by T. C. Koopmans as early as 1948, linear programs can be used to approximate the broad class of convex functions commonly encountered in economic planning.

Linear programs may be viewed as a generalization of the Leontief Input–Output Model, one important difference being that alternative production processes (activities) are allowed to compete; another being the representation of capacity as an input that becomes available at a later point in time as an output (possibly depreciated). Solving a model with alternative activities requires not only software for efficiently solving on computers large systems of equations as in the Leontief case, but also software for selecting the best combination from an astronomical number of possible combinations of activities. (See the entry simplex method for solving linear programs.)

Formulating a Linear Program

Finding an optimal product mix (for example blend of gasoline, or metals, or mix of nuts, or animal feeds) is a typical application. For example, a manufacturer wishes to purchase at minimum total cost a number of solder alloys *A*, *B*, *C*, *D* which are available in the market-place in order to melt them down to make a blend of 30% lead, 30% zinc, and 40% tin. Their respective costs per pound are shown in Table 1.

Suppose 100 pounds of blend is desired and X_A, X_B, X_C, X_D are the unknown number of pounds of *A, B, C, D* to be purchased. The problem to be solved is clearly: find Z and (X_A, X_B, X_C, X_D) ≥ 0, such that:

$$\begin{aligned} 0.1X_A + 0.1X_B + 0.4X_C + 0.6X_D &= 30 \\ 0.1X_A + 0.3X_B + 0.5X_C + 0.3X_D &= 30 \\ 0.8X_A + 0.6X_B + 0.1X_C + 0.1X_D &= 40 \\ \hline 4.1X_A + 4.3X_B + 5.8X_C + 6.0X_D &= Z(\min) \end{aligned}$$

This example can be solved in a few seconds on a personal computer.

Linear Programming, Table 1

	Alloy				
Composition	A	B	C	D	Desired blend
% Lead	10	10	40	60	30%
% Zinc	10	30	50	30	30%
% Tin	80	60	10	10	40%
Cost/Ib	4.1	4.3	5.8	6.0	Minimize cost per pound

The standard form of a linear program is: find min z, $x = (x_1,...,x_n) \geq 0$:

$$Ax = b, \quad cx = z(\min)$$

where A is a m by n matrix, b a column vector of m components and c a row vector of n components. The matrix A of coefficients is referred to as the *technology matrix.*

One way to formulate a linear program is to begin by (*a*) listing various constraints such as resources availability, demand for various goods by consumers, known bounds on productive capacity; (*b*) listing variables to be determined representing the levels of activities whose net inputs and outputs must satisfy the constraints, and finally (*c*) tabulating the coefficients of the various inequalities and equations.

Since linear programming models can be very large systems with thousands of inequalities and variables, it is necessary to use a special software, called *matrix generators*, to facilitate the model building process. Such systems have millions of coefficients, fortunately most of them are zero. Matrices with very few nonzero elements are called *sparse.* The World Bank uses software called GAMS to generate moderate-size sparse matrices A by rows. Another type of software called OMNI has been developed by Haverly Systems and has been used to generate very large sparse matrices by columns. When a model is formulated by columns, it is called *Activity Analysis*: the column of coefficients of a variable is the same as a recipe in a cook book – these are the input and output flows required to carry out one unit of an activity (or process). The variables, usually non-negative, are the unknown levels of activity to be determined. For example the activity of 'putting one unit (pound) of solder alloy A in the blend' has an input of $4.10 and outputs to the blend of 0.10 lb of lead, 0.10 lb of zinc, 0.80 lb of tin.

In economic applications, *output* coefficients are typically stated with + signs and *input* coefficients with – signs. Under this convention, the signs of the coefficients of the Z equation in the blending example should be reversed and, net revenues, $(-Z)$ maximized. In practice, instead of equations in non-negative variables, there can be a mix of equations and inequalities. Simple algebraic steps allow one to pass from one form of the system to another.

L

Primal and Dual Statements of the Linear Program

John von Neumann in 1947 was the first to point out that associated with a linear program is another called its *Dual*, formed by transposing the matrix A and interchanging the role of the RHS b and the 'cost' vector c. The original problem is called the *Primal*. Von Neumann expressed both of these LP in inequality form:

Primal: $\min \ \overline{z} = cX : AX \geq b, X \geq 0, \quad (P)$

Dual: $\min \ z = Y'b : Y'A \leq c, \quad Y \geq 0, \ (D).$

where Y' is the transpose of column vector Y.

If we denote the jth column of A by $A\ (*, j)$, the n inequalities $Y'A \leq c$ may be rewritten as $Y'A\ (^*, j) \leq c_j$ for $j = 1,\ldots,$ n.

(P) expresses the physical constraints of the system under study. The variables Y of the dual (D) can be interpreted as *prices*. Mathematicians call them *Lagrange multipliers.* The dual conditions, $Y'A(^*, j) - c_j \leq 0$ for $j = 1,...,$ n, may appear strange and just the opposite from what one would expect. They state that levels X_j of all activities j that show profit in the economy will

rise to the point that all 'price out' nonprofitable. It turns out that when the value $z = Y'b$ in (D) is maximized, all activities j that are operated at *positive* levels will just *break even,* i.e., just 'clear their books' and that all activities j operating at a strict loss will be operating at zero levels.

The famous duality theorem of von Neumann states that, when there exist 'feasible' solutions $AX \geq b$, $X \geq 0$ to (P) and $Y'A \leq c$, $Y \geq 0$ to (D),

$$\text{Max } \overline{z} = \text{Min } \overline{z}$$

It is easy to prove that *any* feasible solutions to (P) and (D) not necessarily optimum satisfy.

$$\overline{z} = Y'b \leq cX = \overline{z},$$

so that if it happens that $Y^*b = cX^*$, for some feasible $X = X^*$, $Y = Y^*$ then by the duality theorem we know that such a pair (X*, Y^*) are optimal solutions to (P) and (D).

This makes it possible to combine the primal and dual problems into the single problem of finding a feasible solution to the following: find $(X, \overline{X}, Y, \overline{Y}, \theta) \geq 0$:

$$\begin{bmatrix} 0 & A & -b \\ -A' & 0 & c' \\ b' & -c & 0 \end{bmatrix} \begin{bmatrix} Y \\ X \\ 1 \end{bmatrix} = \begin{bmatrix} \overline{Y} \overline{X} \theta \end{bmatrix} \quad \text{(P, D)}$$

where we have introduced two *slack* vectors $\overline{Y} \geq 0$ and $\overline{X} \geq 0$ which turn the inequality relations (P) and (D) into equality relations $AX - \overline{Y} = b$ and $Y'A + \overline{X}' = c$. The last relation is the single equation $Y'b - cX = \theta$ where $\theta \geq 0$ is a scalar.

It we multiply (P, D) by the vector (Y', X', 1) on the left and perform all the matrix multiplications, everything on the left side cancels out because of the skew symmetry of the matrix and we are left with

$$0 = \left[Y' X' 1 \right] \left[\overline{Y} \overline{X} \theta \right]' = \sum_i Y_i \overline{Y} + \sum_j X_j \overline{X}_j + \theta.$$

Because all terms are non-negative, it follows that

$$X_j \overline{X}_j = 0, \quad Y_i \overline{Y}_i = 0, \quad \theta = 0 \quad \text{for all } i \text{ and } j.$$

These are called *complementary slackness* or *Kuhn–Tucker* conditions for optimality.

Zero-Sum Matrix Games

These games can be formulated as a special class of linear programs. The 'row' player chooses row i of a matrix while his opponent, the 'column' player, simultaneously chooses column j. Column player wins an amount a_{ij} if $a_{ij} \geq 0$ otherwise he pays the other player $-a_{ij}$. The *payoff* matrix is $A = [a_{ij}]$. It is called a *zero-sum* game because the sum of the payments each player receives adds up to zero. Von Neumann analysed this game in 1928 and introduced the notion of a *mixed strategy* (Y_1, Y_2, ... ,Y_m), (X_1, X_2, ... ,X_n) which are the probabilities of the players choosing any particular row and column. He showed that there exist optimal mixed strategies, $Y = Y^*$ for the row player and $X = X^*$ for the column player, such that if a player's mixed strategy is discovered by his opponent, it will have no effect on his expected payoff and hence no effect on the expected payoff of his opponent which is the negative of his.

The column player, if he plays conservatively and assumes his mixed strategy will become known to his opponent, will choose his probabilities $X_j \geq 0$ so as to maximize L where Max L and $X \geq 0$ are chosen so that

$$\sum_j a_{ij} X_j \geq L, \quad X_j \geq 0, \quad i = 1, \ldots, m \sum_j X_j = 1. \qquad \text{(C)}$$

Likewise the row player's-optimal mixed strategy, if he plays conservatively, will choose his probabilities $Y_i \geq 0$ so as to minimize K where Min K and $Y \geq 0$ are chosen so that

$$\sum_i Y_i a_{ij} \leq K, \quad Y_i \geq 0, \quad j = 1, \ldots..., n \sum_i Y_i = 1. \qquad \text{(R)}$$

It is not difficult to prove that (C) and (R) are feasible linear programs and each is the dual of the

other. Let $(Y_1^*, \ldots, Y_m^*)$ and $(X_1^*, \ldots, X_m^*)$ be optimal solutions to (R) and (C). Applying the duality theorem, we obtain von Neumann's famous *minimax theorem* for zero-sum bimatrix games:

$$\max\ L = \min\ K = \sum_i \sum_j Y_i^* a_{ij} X_j^*,$$

the expected payoff to the column player.

Decomposition Principle

Linear programming can be used in an iterative mode to aid a Central Authority to allocate scarce resources to factories in an optimal way without having to have detailed knowledge about each factory. Specifically the Central Authority proposes prices on the scarce commodities that induce the factories to submit a summary plan for approval of their requirement for scarce resources. The Central Authority blends these proposed plans with earlier ones submitted and uses them to generate new proposed prices. The entire cycle is then iterated. This method, first proposed by Dantzig and Wolfe in 1960, is known as the D–W or *Primal Decomposition Principle.*

The dual form of the Decomposition Principle is known as Benders Decomposition and was proposed by Benders in 1962. We illustrate it here in the context of a two-period planning problem.

$$\begin{aligned} \text{find}\ \min\ Z = c_1X_1 + c_2X_2 \quad \text{subject to:} \\ b_1 = A_1X_1 \quad (X_1, X_2) \geq 0, \\ b_2 = -B_1X_1 + A_2X_2 \end{aligned}$$

where A_1, B_1, A_2 are matrices, b_1, b_2, c_1, c_2 vectors and $X_t \geq 0$ are the vectors of activity levels to be determined in periods $t = 1$ and 2.

The first period planners determine a feasible plan (p) that satisfies $b = AX_l^p, X_l^p \geq 0$ (augmented by certain necessary conditions, called 'cuts'), which they submit to the second period planners in the form of a vector $B_1X_l^p$ which is used by them to solve the second period *sub* problem:

$$A_2X_2 = b_2 + B_1X_1^p, X_2 \geq 0 c_2X_2 = Z_2(\min).$$

The second period planners respond with a vector of *optimality prices* π_2^k corresponding to the second period if the sub-problem is feasible, or with *infeasibility prices* σ_2^l(obtained at the end of phase 1 of the simplex method) if it is infeasible.

The first period planners then iteratively resolve, their problem augmented by $k' + l'$ additional necessary conditions (cuts) shown below:

$$\begin{aligned} &\text{Find } c_1X_1 + \theta = Z(\min) \\ &\qquad A_1X_1 = b_1, \quad X \geq 0, \\ &\text{optimality cuts}: -(\pi_2^k B_1)X_1 + \theta \geq \pi_2^k b_2,\ k = 1, \ldots, k' \\ &\text{infeasibility cuts}: -(\sigma_2^l B_1)X_1 \geq \sigma_2^l b\ l = 1, \ldots, l' \end{aligned}$$

where $\theta = (c_2X_2)$ is treated as an unknown variable. The interative process stops if $\theta = Z_2$, or $Z_2 - \theta = \Delta > 0$ is small enough.

Note that the additional conditions imposed on Period 1 are expressed in terms of Period-1 variables and θ only. These serve as surrogates for future periods (in this example for only one future period). The decomposition principle allows one to solve a multi-time-period problem one period at time and pass the ending conditions of one period on to initiate the next and to pass back price vectors to earlier periods that are translated into policy constraints called cuts. Applying this same approach to a multi-stage production line, one obtains an iterative process that can be viewed as an intelligent control system with learning.

See Also

- ▶ Efficient Allocation
- ▶ Non-linear Programming
- ▶ Simplex Method for Solving Linear Programs

Bibliography

Beale, E.M.L. 1954. Linear programming by the method of leading variables. Report of the Conference on Linear Programming, May, arranged by Ferranti Ltd, London.

Benders, J.F. 1962. Partitioning procedures for solving mixed-variable programming problems. *Numerische Mathematik* 4: 238–252.
Charnes, A., W.W. Cooper, and B. Mellon. 1952. Blending aviation gasolines – A study in programming interdependent activities in an integrated oil company. *Econometrica* 20 (2): 135–159.
Dantzig, G.B. 1948. *Programming in a linear structure*. Washington, DC: Comptroller, USAF.
Dantzig, G. 1949, 1951. Programming of interdependent activities, II, mathematical model. In *Activity analysis of production and allocation,* ed. T.C. Koopmans, 330–335. New York: Wiley; also published in *Econometrica* 17(3 and 4), 1949, 200–211.
Dantzig, G.B. 1963. *Linear programming and extensions*. Princeton: Princeton University Press.
Dantzig, G.B., and P. Wolfe. 1960. A decomposition principle for linear programs. *Operations Research* 8 (1): 101–111.
Kantorovich, L.V. 1939. *Mathematical methods in the organization and planning of production*. Publication House of the Leningrad State University. Translated in Management Science 6 (1960), 366–422.
Koopmans, T.C., ed. 1951. *Activity analysis of production and allocation*. New York: Wiley.
Kuhn, H.W. and Tucker, A.W. 1951. [The symposium was held in 1950, but the proceedings volume was published in 1951.] Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability,* ed. J. Neyman. Berkeley: University of California Press, 481–492; also in *Econometrica* 19(1) (1951), 50–51 (abstract).
Leontief, W. 1951. *The structure of the American economy, 1919–1939*. New York: Oxford University Press.
von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320. Translated by *Sonya Bargmann in contributions to the theory of games*, vol. 4, ed. A.W. Koopmans and R.D. Luce, Annals of Mathematics study no. 40. Princeton: Princeton University Press, 1959, pp. 13–42.
von Neumann, J. 1937. Über ein ökonomisches Gleichungssytem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. Ergebnisse eines Mathematischen Kolloquiums No. 8. Translated in *Review of Economic Studies 13* (1), (1945–46) 1–9.
von Neumann, J. 1947. Discussion of a maximization problem. Manuscript, Institute for Advanced Study. Princeton.
von Neumann, J. 1963. In *Collected works*, ed. A.H. Taub, vol. VI, 89–95. Oxford: Pergamon Press.
Vajda, S. 1956. *The theory of games and linear programming*. New York: Wiley.
Wood, M.K. and Dantzig, G.B. 1949, 1951. The programming of interdependent activities: general discussion. In *Activity analysis of production and allocation,* ed. T.C. Koopmans, 15–81. New York: Wiley, 1951, also in *Econometrica* 17(3 and 4) (1949), 193–199.

Linkages

Antonio Ciccone

Abstract

Economic activities in different industries are linked to each other through aggregate income (horizontal linkages) and input–output relationships (vertical linkages). Could such linkages give rise to vicious circles of underdevelopment or virtuous circles of development when there are increasing returns to scale at the firm level?

Keywords

Cost linkages; Demand linkages; Horizontal and vertical linkages; Increasing returns to scale; Industrialization; Input chains; Linkages; Multiple equilibria; Pre-industrial production methods; Underdevelopment traps

JEL Classifications

O1

Introduction

Economic activities in different industries are linked to each other through aggregate income (horizontal linkages) and input–output relationships (vertical linkages). Could such linkages give rise to vicious circles of underdevelopment or virtuous circles of development when there are increasing returns to scale at the firm level? A standard account of a vicious circle goes as follows. Small-scale production methods in industry *A* lead to low output and income. This translates into low demand for industry *B*, which therefore also ends up using small-scale production methods and generating low output and income. The result is low demand for industry *A*, which justifies the small-scale production methods used in this industry. Low aggregate output and income are seen as the result of a

vicious circle because the same economic environment is thought to be compatible with a high-income equilibrium where all industries use technologies that achieve high productivity at large scale. This high-income equilibrium is sustained by a virtuous circle. Large-scale production methods in industry *A* are profitable because of high income in industry *B*, and vice versa.

We will show that vicious or virtuous circles based on demand linkages are subject to a simple fallacy if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. Still, vertical demand linkages will give rise to vicious or virtuous circles if increasing-returns-to-scale technologies use intermediate inputs more intensively than the technologies they replace. And horizontal demand linkages will do so if firms adopting increasing-returns-to-scale technologies must pay a compensating wage differential. Moreover, when there are both vertical demand and cost linkages, underdevelopment traps can be consistent with economic principles even if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. We first discuss the role of horizontal demand linkages, then that of vertical demand linkages, and finally turn to vertical cost linkages.

Horizontal demand linkages. Imagine an economy populated by households and by firms in different industries. Suppose that each industry sells only to households. Assume also that the amount households spend on each industry is independent of prices (industry demand functions are unit elastic). In this case, demand linkages among industries are said to be *horizontal*. This simply means that economic activity in one industry affects spending on other industries only through the aggregate income of households.

Could horizontal demand linkages lead to economies being trapped into a situation of low income due to a vicious circle of low income and output? Rosenstein-Rodan (1943) and Nurkse (1953) thought so. They imagined a situation where low aggregate income was an obstacle to the adoption of technologies that achieve high productivity at large scale. But large-scale production methods would be profitable if all industries adopted them, because incomes generated in one industry would create demand for other industries.

The elements necessary for underdevelopment traps to be consistent with economic principles have always been subject to debate. Increasing returns to scale appeared to be crucial. But Fleming (1955) made clear that this was not enough. He imagined a situation where, because of low aggregate income, industry *A* cannot make a profit from adopting the increasing-returns-to-scale technology and that the same is true for industry *B*. Is it possible that the increasing-returns-to-scale technology becomes profitable if both *A* and *B* adopt it? Consider forcing *A* to adopt. In this case, the loss made in industry *A* will lower aggregate income. As a result, industry *B* will now face even lower demand and therefore make an even greater loss if it adopts the increasing-returns-to-scale technology. This means that aggregate income will fall further if we also force industry *B* to adopt the increasing-returns-to- scale technology. Hence, if the adoption of increasing-returns-to-scale technologies is unprofitable for any single industry, adoption in all industries will not be profitable either. Increasing returns alone can therefore not explain why industrialization does not take place although it would ultimately be profitable.

All accounts of underdevelopment traps did in fact feature (several) additional elements. In particular, Rosenstein-Rodan maintained that firms using large-scale production methods had to pay a compensating wage differential (partly because of the higher costs of living in urban areas, where industrial firms were located). Section "A Model of Horizontal Demand Linkages" follows Murphy et al. (1989) in showing that underdevelopment traps may emerge when firms adopting the increasing-returns-to- scale technologies must pay a compensating wage premium.

Vertical demand linkages. Suppose now that industries sell goods to households and each other (to be used as intermediate inputs). Economic activity in one industry can then affect demand in another industry even if aggregate income remains

unchanged. As a result, there are said to be *vertical* linkages. For example, consider the situation where industry *B* buys from *A* (industry *A* is *upstream* of *B*). In this case there is a *vertical demand* linkage as demand for the upstream industry *A* will depend on the economic activity in downstream industry *B*. There could also be a *vertical cost* linkage because the cost of production in downstream industry *B* is partly determined by the cost of goods produced in upstream industry *A*.

While the effects of horizontal demand linkages on economic development have always been subject to some controversy, there appears to be a consensus among early contributors that vertical demand linkages can lead to underdevelopment traps when technologies are subject to increasing returns to scale (Fleming 1955; Scitovsky 1954; Hirschman 1958). It is simple to show however that this is not the case if increasing-returns-to-scale technologies differ from pre-industrial-technologies only in that they are more productive at large scale. To see this, note that with vertical demand linkages the adoption of increasing-returns-to-scale technologies affects aggregate income directly and indirectly: directly through the profits made in the adopting industry, and indirectly through the profits made in supplying (upstream) industries. It would therefore seem that increasing-returns-to-scale technologies could be unprofitable in the adopting industry but still increase aggregate income. But this cannot happen when the increasing-returns-to-scale and the pre-industrial technologies use upstream inputs with the same intensity. In this case, the increase in the value of upstream goods demanded by a firm adopting increasing-returns-to-scale technologies is always a fraction of the (absolute value of the) loss that it makes. Moreover, as profits cannot exceed revenues, the increase in profits in supplying industries is necessarily smaller than the increase in the value of goods they sell. It therefore follows that the increase in profits in supplying industries (the positive indirect effect) can never compensate for the loss made in the industry adopting the increasing-returns-to-scale technology.

The empirical evidence indicates that the intermediate-input intensity of production increases with a country's level of industrialization. Increasing-returns-to- scale technologies may therefore be using intermediate inputs more intensively than the production methods they replace. Section "Vertical Demand Linkages in an Input Chain Model" draws on Ciccone's (2002) model of input chains to show that vertical linkages can in this case explain why countries may be trapped into a vicious circle of underdevelopment, and why escaping this trap may be associated with large gains in aggregate income and productivity.

The interplay of vertical cost and demand linkages. The greater demand for intermediate inputs brought about by industrialization (vertical demand linkages) may partly be caused by falling intermediate input prices (vertical cost linkages). Falling intermediate input prices, on the other hand, are possible because of the higher productivity of large-scale production methods. Vertical cost and demand linkages therefore feed on each other (Young 1928; Okuno-Fujiwara 1988; Rodriguez-Clare 1996). For example, Rodriguez-Clare considers a small open economy framework where the entry of new intermediate input varieties lowers the cost of intermediate inputs relative to labour, which leads final-good producers to substitute towards intermediate inputs. When this substitution effect is strong enough, it translates into greater revenues and profits for intermediate-input producers, which may validate intermediate-input producers' decision to start up new varieties in the first place. Rodriguez-Clare shows that this interplay of vertical demand and cost linkages may lead to two equilibria: a low-income equilibrium where final-good producers use labour-intensive production methods because of the limited range of intermediate inputs available, and a high-income equilibrium where a large variety of intermediate inputs leads final-good producers to use intermediate- input intensive production methods. Okuno-Fujiwara (1988) considers a situation where vertical demand and cost linkages interact because greater demand for intermediate inputs leads to lower prices due to competition among a larger number of Cournot oligopolists. The final section of this entry uses the model with input chains to show that the interplay between

vertical demand and cost linkages can result in underdevelopment traps even if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale.

A Model of Horizontal Demand Linkages

We will now examine the role of horizontal demand linkages for economic development using the model of Murphy et al. (1989) (for a historical and methodological perspective on the horizontal-linkages literature, see Krugman 1993, 1994). The first step is to describe the model set-up – the household sector, the production sector, and market structure. The second step is to characterize equilibrium prices and equilibrium allocations.

Households. There are L households, each of whom supplies one unit of labour in elastically (labour is the only production factor in this model and serves as the numeraire). Households spend an equal share of their incomes on each of the N goods produced in the economy.

Production. Each of the N goods demanded by households can be produced using two different production methods: a *pre-industrial* method requiring one unit of labour for each unit of output produced, and an *industrial* or increasing-returns-to- scale method, which is more efficient at the margin but subject to a fixed labour requirement (f). Formally, the increasing-returns-to-scale production method requires

$$l_i = f + cq_i \tag{1}$$

units of labour to produce qi units of good i, where $f > 0$ and $1 > c > 0$.

Industry wage premium. Working in the industrial sector generates a disutility $v \geq 0$ for households. Hence, relative to pre-industrial firms, industrial firms will have to pay a wage premium $v \geq 0$ as a compensating wage differential.

Market structure. Many firms are assumed to know the pre-industrial method to produce good i. As a result, the pre-industrial sector (also called competitive fringe) will be characterized by perfect competition. By contrast, only a single firm is taken to have the ability to produce each good in the industrial sector. These firms set prices optimally, taking the prices of all other firms as given. The labour market is taken to be perfectly competitive.

What keeps this model simple to analyse is that the equilibrium price of each good is unity whether the good is produced by the pre-industrial or the industrial sector. To see this, note that perfect competition and constant returns to scale in the pre-industrial sector imply that the price of goods produced in this sector must be equal to unity. A higher price would mean strictly positive profits and therefore further entry of pre-industrial producers, while a lower price would mean that no pre- industrial producer could break even. Now consider goods produced in the industrial sector. Clearly, the industrial producer will not set a price above unity, as she would lose the entire market to pre-industrial producers in this case. Moreover, industrial producers do not have an incentive to set a price below unity either, as households spend the same fraction of income on their good irrespectively of the price. Hence, industrial producers find it optimal to use a limit pricing strategy, setting prices exactly equal to the marginal cost of pre-industrial producers. As a result, the price of each of the N goods is equal to unity independently of the production method.

Pre-industrial equilibrium. Under what conditions will there be an equilibrium where all goods are produced with the pre-industrial method? In such an equilibrium, firms just break even, and aggregate income Y in the economy is therefore equal to aggregate labour income L. Because households spread income equally among all N goods, the quantity of good i demanded and supplied is $q_i = L/N$. The remaining question is whether firms in the industrial sector have an incentive to adopt the increasing-returns-to-scale method. The potential profit of such firms is $\pi_i = q_i^m - \left(f + cq_i^m\right)(1 + v)$, where q_i^m is the demand faced by the industrial producer of good i. As industrial and pre-industrial producers set the same price, the first industrial producer faces exactly the same demand as the pre-industrial producers she replaces, $q_i^m = L/N$. Her profits are therefore

$$\pi_i = L/N - (f + cL/N)(1 + v). \quad (2)$$

If $\pi_i < 0$, an industrial producer has no incentive to adopt the increasing-returns-to- scale method, and it will be an equilibrium for all goods to be produced with the pre- industrial method. Hence, (2) implies that there is an equilibrium where all goods are produced with the pre-industrial method if

$$L(1 - c(1 + v)) < F(1 + v), \quad (3)$$

where $F \equiv fN$.

Industrial equilibrium. What about equilibria where all goods are produced using the industrial method? We already know that prices of all goods will be equal to unity in this equilibrium also. Moreover, households will keep spending the same share of income on all goods. Hence, all industries will employ the same amount of labour, L/N, in equilibrium. (1) therefore implies that the value of production in each industry is $(L/N - f)/c$. Summing across the N industries in the economy yields a value for gross domestic product, and hence aggregate household income, of $Y = (L - F)/c$ (recall that $F \equiv fN$).

Do firms make the profit necessary to sustain the industrial production method when all production takes place in the industrial sector? Profits of firms in the industrial sector are $\pi_i = q_i^m - \left(f + cq_i^m\right)\ (1 + v) \geq 0$, where $q_i^m$is the demand faced by the industrial producer of good i, $q_i^m = Y/N(L - F)/cN$. Hence, there will be an equilibrium where firms using the increasing-returns-to-scale method make a profit if

$$L(1 - c(1 + v)) \geq F. \quad (4)$$

Efficient allocation. When is the adoption of increasing-returns-to-scale technologies efficient? The aggregate value of production is $Y = (L - F)/c$ when industrial production methods are used and $Y = L$with pre-industrial methods. The amount of goods necessary to pay the compensating wage differential when all workers are employed in the industrial sector is vL. Hence aggregate welfare will be higher with industrial production methods if and only if $(L - F)/c - vL \geq L$, or

$$L(1 - c(1 + v)) \geq F. \quad (5)$$

Note that (4) and (5) coincide. Hence, an industrial equilibrium exists if and only if it is efficient.

Multiple equilibria and underdevelopment traps. Only one of the two inequalities in (3) and (4) can hold if there is no industry wage premium ($v = 0$).Hence, the equilibrium is unique in this case and, as a result, there cannot be development traps. Moreover, because an industrial equilibrium exists if and only if it is efficient, economies in a pre-industrial equilibrium actually do the best they can given the economic environment.

But when there is an industry wage premium ($v > 0$) there may be multiple equilibria as the inequalities in (3) and (4) can both be satisfied. When this is the case, economies may be stuck in a pre-industrial equilibrium, although the same economic environment would be compatible with an (efficient) industrial equilibrium. To understand why, suppose the economy is in a pre-industrial equilibrium when we force an industry to adopt the increasing-returns-to-scale technology. If (3) holds, then the adopting firm will make a loss. Still, its contribution to aggregate income is strictly positive. To see this, note that demand for this industry is L/N, and that this is also the amount of labour required to produce the amount of goods demanded using the pre-industrial production methods. Production with the increasing-returns-to-scale technology requires $cL/N + f$ units of labour, which is strictly smaller than L/N if (4) holds. Hence, the adoption of the increasing-returns-to-scale technology saves labour in the adopting industry, and therefore increases aggregate output and income. This increases demand faced by other industries and therefore raises the profitability of further adoption of the increasing-returns-to-scaletechnology. Eventually,industrialization raises aggregate income enough for increasing-returns-to-scale industries to break even. Hence, the industrial equilibrium can be seen as the result of a virtuous circle. The adoption of increasing-returns-to-scale technologies raises aggregate income and therefore the profitability of adopting increasing-returns-to-scale technologies. At the same time, the economic environment also allows for a

development trap where low aggregate income is both the cause and the consequence of the failure to adopt increasing- returns-to-scale technologies.

Vertical Demand Linkages in an Input Chain Model

The economic activity of different industries is linked to each other because the output of some industries is used as input in other industries. Can such vertical linkages give rise to vicious circles of underdevelopment or virtuous circles of development when there are increasing returns at the firm level? We will show that – just as for horizontal linkages – this cannot happen if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale.

Chenery et al. (1986) comparative study of industrialization shows, however, that the industrialization of countries has typically been accompanied by an increase in the intermediate-input intensity of production. This suggests that industrial technologies may use intermediate inputs more intensively than the technologies they replace. We will therefore start by analysing a model of development where increasing-returns-to-scale technologies use intermediate inputs more intensively than pre-industrial technologies.

It will be useful to analyze the consequences of vertical linkages for industrialization in a framework that is as close as possible to the model of horizontal linkages of Murphy, Shleifer and Vishny. In particular, the aggregate amount of labour supplied by households continues to be L and households spend an equal share of their incomes on each of the N goods produced in the economy. On the production side, we continue to assume that each good can be produced using two different production methods, namely, a pre-industrial method and an industrial (increasing-returns-to-scale) method. The pre-industrial method requires one unit of labour for each unit of output. The increasing-returns-to-scale method will turn out to be cheaper at the margin but subject to a fixed labour requirement f. Many firms know the pre-industrial method, but for each good there is only a single firm with the ability to produce in the industrial sector.

Input chains and industrial production. The key difference with the horizontal linkages model is that now the increasing-returns-to-scale method is taken to be more intermediate-input intensive than the pre-industrial method. One way to model the intermediate-input structure of the economy is to think of goods being produced in S different locations along a river. Each location produces H different goods (the total number of goods is $N \equiv HS$). Goods at location 1 are produced using labour only. Goods at any location $s > 1$, on the other hand, are produced using all goods at location $s-1$. This implies that all goods at locations $s < S$ may face intermediate-input demand from downstream industries in addition to consumption-goods demand from households (the exception are the H goods furthest downstream, at location S, which face consumption-goods demand only). In particular, we assume that, after having incurred the overhead labour cost, one unit of any good j located at $s > 1$ can be produced with c units of an intermediate-input composite $z_{j,s}$ that combines all H goods produced at location $s-1$,

$$z_{j,s} = \prod_{i=1}^{H} \left(H q_{i,s-1}\right)^{1/H}, \tag{6}$$

where $q_{i,\,s-1}$ is the input of good i at location $s-1$.This formulation implies that industrial firms spend the same amount on all upstream inputs. As a result, the marginal cost of the intermediate-input composite necessary for industrial production at location $s > 1$ is simply a geometrically weighted average of prices $p_{i;s-1}$ of the H up stream goods,

$$MC_s = \prod_{i=1}^{H} p_{i,s-1}^{\frac{1}{H}}. \tag{7}$$

Industrial production for goods at location $s = 1$ requires f units of overhead labour and c units of labour for each unit of output. (The assumption that the industrial overhead requires labour only while production at the margin requires

intermediate inputs only simplifies the analysis considerably. Ciccone (2002) analyses the case where production of the overhead and at the margin use both labour and intermediate inputs.)

Just as in the horizontal linkages model, industrial firms find it optimal to use a limit pricing strategy for consumption goods vis-à-vis the competitive fringe. Their intermediate-input pricing strategy is potentially more complicated but also simplifies to a limit pricing strategy vis-à-vis the competitive fringe when H is sufficiently large.

Pre-industrial equilibrium. When will there be an equilibrium where all goods are produced with the pre-industrial method? It turns out that if H is sufficiently large the condition is

$$L(1-c) < F, \tag{8}$$

which coincides with the condition for a pre-industrial equilibrium in the Murphy, Shleifer, and Vishny model of horizontal linkages. To see this, suppose that all goods are produced with the pre-industrial technology and their price is unity. When (8) holds, any single firm adopting the increasing-returns-to-scale method to produce consumption goods will make a loss. Moreover, when H is sufficiently large, (7) also implies that single industrial firms are unable to generate intermediate-input demand for their good even if they lower their price to the marginal cost of production. To see this, suppose that one industrial firm at location $S-1$ is considering selling its good at marginal cost to firms at location S in order to generate intermediate-input demand. In this case, one of the H inputs of potential industrial firms at S would become available at price c and (7) implies that the marginal cost of production would therefore fall from c to $c^{(1+H)/H}$ (recall that the remaining $H-1$ inputs are available at price of unity). Goods at S face demand L/N, which comes exclusively from households as there are no upstream industries. Hence, profits of the potential industrial firm at S producing at marginal cost $c^{(1+H)/H}$ would be $(1-c^{(1+H)/H})L/N-f$, which is strictly negative if (8) holds and H is large enough. Potential industrial firms at location S would therefore find it unprofitable to start production even after the price cut, which implies that potential industrial firms at location $S-1$ must break even on consumption-goods demand only. Applying the same argument sequentially to potential industrial firms in locations $S-2$, $S-3$, ..., 1 yields that pre-industrial production of all goods is an equilibrium when (7) holds and H is sufficiently large.

Industrial equilibrium. To determine the conditions for the existence of an industrial equilibrium, it is necessary to determine aggregate income when all goods at location σ and upstream of location σ are produced with the increasing-returns-to- scale technology. This turns out to be straightforward. If aggregate income is Y, the quantity of each good demanded by households is Y/N. The intermediate-input structure implies that industrial production of Y/N units of each of the H goods at location σ requires cY/N units of each of the H goods at location $\sigma-1$.Hence, as Y/N units of good $\sigma-1$ are demanded by households, production of each good at $\sigma-1$ must be $Y/N+cY/N$. Production of this quantity of goods at $\sigma-1$ requires $C(Y/N+cY/N)$ units of each good at $\sigma-2$. Adding the Y/N units of goods at $\sigma-2$-demanded by households, yields that production at $\sigma-2$ must be $Y/N+cY/N+c^2Y/N$. Continuing all the way up stream yields that the total production of each of the H goods at location 1 must be

$$\begin{aligned} q_1 &= Y/N + cY/N + c^2Y/N + \ldots \\ &\quad + c^{\sigma-1}Y/N \\ &= \frac{1-c^{\sigma}}{1-c}Y/N. \end{aligned} \tag{9}$$

To turn to the labour market, f units of labour must be used as overhead in the production of each good produced with the industrial technology. Moreover, Y/N units of labour are required for the production of each good produced with the pre- industrial technology. Hence, the amount of labour available for production at the margin of the H goods at $s=1$ is $L-\sigma Hf-(N-\sigma H)Y/N$. Labour market clearing requires $cHq_1 = L-\sigma Hf-(N-\sigma H)Y/N$. Substituting (9) yields aggregate income in an economy where the σ industries furthest upstream have industrialized:

$$Y(\sigma) = \frac{L - F(\sigma H/N)}{c\theta[\sigma](\sigma H/N) + (1 - (\sigma H/N))} = \frac{L - F(\sigma H/N)}{1 - (\sigma H/N)(1 - c\theta[\sigma])}, \quad (10)$$

where

$$\theta[\sigma] \equiv \frac{1 - c^{\sigma}}{(1 - c)\sigma}.$$

$c\theta[\sigma]$ has a simple interpretation. It is the amount of labour required to produce one additional unit of goods located at σ if all industries upstream of (including) σ have adopted the industrial technology. Note that the amount of labour required to produce one additional unit of goods at location σ falls the longer the industrial input chain ($\theta[\sigma]$is strictly decreasing in σ).

The intermediate-input structure implies that the demand for goods is greater the further upstream they are located. Hence, profits from adopting the increasing- returns-to-scale technology fall the further downstream industries are located. An equilibrium where all industrial firms make a profit will therefore exist if goods produced furthest downstream (at location S) can be produced using the increasing- returns-to-scale technology without a loss. Because firms furthest downstream sell to households only, their sales are equal to aggregate income divided by the number of goods, $Y[S]/N$ (recall that all firms set prices optimally at unity). As a result, their profits are positive if and only if $\pi_S = (1 - c)(Y[S]/N) - f \geq 0$ or, to make use of (10),

$$(1 - c)L \geq (c\theta[S] + (1 - c))F. \quad (11)$$

Multiple equilibria and underdevelopment traps. Comparison of (8) and (11) yields that, with input chains ($S > 1$), it is possible for the pre-industrial equilibrium and the industrial equilibrium to exist side by side. (When $S = 1$ then $\theta = 1$ and the model is that of Murphy, Shleifer, and Vishny without an industry wage premium.) This is because the adoption of increasing-returns-to-scale technologies now has a direct and indirect effect on income. The direct effect is given by the profit or loss in the adopting industry. The indirect effect is equal to the profits generated upstream of the adopting industry. When the indirect profits generated by the increased intermediate-input demand more than offset direct losses of industrial technologies, then industrialization increases aggregate income. As a result, further industrialization becomes more profitable. When (7) and (10) hold simultaneously, this effect is strong enough to ensure that all industrial firms make a profit once all goods are produced with increasing-returns-to-scale technologies.

The pre-industrial and industrial equilibrium can exist side by side even if aggregate income is much greater in the industrial equilibrium. Note that aggregate income in the industrial equilibrium is $Y[S] = (L - F)/c\theta[S]$, see (10). As intermediate-input chains become longer, $\theta[S]$ in (10) tends to zero, and aggregate income in the industrial equilibrium increases. Aggregate income in the pre-industrial equilibrium, on the other hand, is independent of S as production does not rely on intermediate inputs. Moreover, the range of parameter values for which the industrial equilibrium exists increases. Hence, long input chains imply that equilibrium multiplicity is more likely and also that the aggregate income difference between industrial and pre-industrial equilibria may be very large.

Vertical linkages and equilibrium uniqueness. To see that the equilibrium is unique when increasing-returns-to-scale technologies use intermediate inputs as intensively as pre-industrial technologies, note that costs of production plus profit must add up to the value of firms' sales, $COST + \pi = q$. Suppose that intermediate inputs are a share α of costs of production for both the pre-industrial and the industrial production method. In this case, the demand for goods produced at $s - 1$ is equal to $\alpha COST_S = \alpha(q_s - \pi_s)$. Now suppose that all goods upstream of σ are produced with the increasing-returns-to-scale technology. Is it possible that aggregate income increases with the adoption of the increasing-returns-to-scale technology at σ even if the adopting firm makes a loss? A switch to industrial production at σ does not affect the value of goods

produced at this location (q_σ is unchanged). Hence, the adoption of the increasing-returns-to-scale technology at σ increases demand for each good produced at $\sigma - 1$ by $\alpha\pi_\sigma/H$. Loss-making industrialization at σ therefore leads to greater demand at $\sigma - 1$. But the profits generated by this input demand can never be greater than the initial loss π_σ. To see this, notice that total profits at location $\sigma - 1$ increase by $-(1 - c)\alpha\pi_\sigma$. Total profits at $\sigma - 2$ increase by $-(1 - c)\alpha^2 c\pi_\sigma$,where $-\alpha^2 c\pi_\sigma/H$ is the increase in demand for each good produced at $\sigma - 2$. The general formula is that total profits at location $\sigma - i$ increase by $-(1 - c)\alpha^i c^{i-1}\pi_\sigma$. Summing profits across all locations yields $-(1 - c)\pi_\sigma\alpha[1 + \alpha c + (\alpha c)^2 + \cdots + (\alpha c)^{\sigma - 1}]$, which is smaller than $-(1 - c)\pi_\sigma\alpha[1 + \alpha c + (\alpha c)^2 + \cdots] = -\pi_\sigma(\alpha - \alpha c)/(1 - \alpha c)$. Hence, $\alpha \leq 1$ implies that the sum of profits generated upstream of s by loss-making industrialization at s is always smaller than the initial loss (π_σ). Loss-making industrialization necessarily lowers aggregate income. The aggregate demand externality necessary for multiple equilibria is therefore absent when increasing- returns-to-scale technologies are no more intermediate-input intensive than pre-industrial technologies.

Vertical Demand and Cost Linkages with Input Chains

So far firms adopting increasing returns to scale technologies did not have an incentive to cut prices. This eliminated virtuous circles of development where lower intermediate-input prices (vertical cost linkages) and greater intermediate-input demand (vertical demand linkages) feed on each other. A simple way to capture the interplay between vertical demand and cost linkages is to suppose that firms in the competitive fringe can produce one unit of goods at location $s > 1$ with $1 + \varepsilon > 1$ units of the intermediate-input composite in (6) or one unit of labour. That is, firms have access to two modes of production, a labour-intensive mode and an intermediate-input intensive mode. The exception continues to be goods at location 1, for which there is a labour-intensive mode of production only. Industrial firms at locations $s > 1$ also have access to a labour-intensive and an intermediate-input intensive mode of production, but are more efficient than pre-industrial firms at the margin. Once they have incurred the overhead labour requirement f, industrial firms can produce one unit of output with $c(1 + \varepsilon) < 1$ of the intermediate-input composite in (6) or $c < 1$ units of labour. Industrial firms producing goods at location 1 have access to the labour-intensive mode of production only. The assumption that the overhead is produced using labour only continues to simplify the analysis considerably. A new by-product of this assumption is that industrial firms now actually use intermediate inputs less intensively than pre-industrial firms at the same factor prices – the opposite of what we assumed in the previous section.

Pre-industrial equilibrium with labour-intensive production. Can there be an equilibrium where all goods are produced with the pre-industrial technology using labour only? The marginal cost of production with the pre-industrial technology in the labour-intensive mode is unity. Hence, the price of all goods would be equal to unity. To see that these prices make it optimal to use the labour-intensive mode of production, note that they imply that the marginal cost of intermediate-input composites in (7) is unity. The marginal cost of production using the intermediate- input intensive mode compared with the labour-intensive mode is therefore $1 + \varepsilon > 1$ (in the pre-industrial as well as the industrial sector). Hence, all firms will find it optimal to use the labour-intensive mode of production.

In a pre-industrial equilibrium, the adoption of the increasing-returns-to-scale technology by a single firm must lead to losses. If industrial firms can count on consumption-goods demand only, this will be the case if $L(1 - c) < F$. But an industrial firm may be able to generate additional demand by getting industries just downstream to switch to an intermediate-input intensive mode of production. While this can happen in principle, it will not happen if H is sufficiently large. To see this, consider the case where a single industrial firm supplies its good to downstream industries at

marginal cost. In this case, (7) yields that the marginal cost of the intermediate input-intensive mode of production relative to the labour-intensive mode becomes $c^{1/H}(1+\varepsilon)$, which will be greater than unity when H is sufficiently large (recall that $1+\varepsilon > 1$), Hence, a single industrial firm cannot generate downstream intermediate-input demand even if it reduces its price to marginal cost. For H sufficiently large, a pre-industrial labour-intensive equilibrium will therefore exist if $L(1-c) < F$.

Industrial equilibrium with intermediate-input intensive production. When is there an industrial equilibrium where all firms use the intermediate-input intensive mode of production? To simplify the analysis, suppose that industrial firms can price discriminate between households and industrial users of their goods. As before, industrial firms will find it optimal to follow a limit pricing strategy when it comes to sales to households. Industrial firms will therefore price consumption goods at unity. When it comes to intermediate-input sales to downstream industries, industrial firms must also take into account that users will switch to the labour-intensive mode of production if the cost of the intermediate-input composite is greater than $1/(1+\varepsilon)$. Hence, each industrial firm will find it optimal to set a limit price of $1/(1+\varepsilon)$. for intermediate inputs if other industrial intermediate-input suppliers do the same.

Aggregate income in the industrial equilibrium where all firms use the intermediate-input intensive mode of production can be determined following the argument that led to (10). The only difference is that an additional unit of all goods at location $s > 1$ now translates into a demand of $c(1+\varepsilon)$units of each good at location $s-1$. Aggregate income when all goods are produced with the industrial technology in the intermediate-input intensive mode is therefore $Y[S] = (L-F)/c\hat{\theta}[S]$ where

$$\hat{\theta}[S] \equiv \frac{1-(c(1+\varepsilon))^s}{(1-(c(1+\varepsilon))^S}. \tag{12}$$

An industrial equilibrium exists if the firm furthest downstream can break even given the demand for consumption goods, $\pi_s = (1-c)(1+\varepsilon))(Y[S]/N - f \geq 0$ or, to make use of the expression for aggregate income just above, $(1-c(1+\varepsilon))L \geq \Big((c\hat{\theta}[S] + (1-c(1+\varepsilon))\Big)F$.

Multiple equilibria with vertical demand and cost linkages. There will be multiple equilibria if both $L(1-c) < F$ and $(1-c(1+\varepsilon))L \geq \Big((c\hat{\theta}[S] + (1-c(1+\varepsilon))\Big)F$. This implies that the pre-industrial equilibrium with labour-intensive production and the industrial equilibrium with intermediate-input intensive production may exist side by side if and only if there are input chains ($\hat{\theta}[S] < 1$). The virtuous circle sustaining industrial equilibria now consists of an interplay between vertical demand and cost linkages. The increase in the intermediate-input intensity of production necessary for increasing-returns-to- scale technologies to be profitable (vertical demand linkages) comes about because the adoption of increasing-returns-to-scale technologies translates into falling intermediate-input prices (vertical cost linkages). Note that, for this virtuous circle to be operative, the elasticity of substitution between intermediate inputs and labour in industrial production must be greater than unity (our model assumed that this elasticity is infinity for simplicity). In a pre-industrial equilibrium, on the other hand, pre-industrial technologies are both the cause and the consequence of labour-intensive modes of production.

Conclusion

Neither horizontal nor vertical demand linkages across industries lead to underdevelopment traps if increasing-returns-to-scale technologies differ from pre-industrial technologies only in that they are more productive at large scale. Nevertheless, theories of underdevelopment based on vicious circles of low demand and low productivity are consistent with economic principles. For example, in the case of vertical demand linkages, there can be development traps if increasing- returns-to-scale technologies use intermediate inputs more intensively than the technologies they replace. More generally, multiple equilibria in our models

exist under assumptions that do not appear to be in contradiction by empirical evidence. The exception is that all our model economies were taken to be closed to international trade, but we could have assumed instead that only some goods are nontradable or that all goods are tradable at some cost (for example, Okuno-Fujiwara 1988; Rodriguez-Clare 1996; Krugman and Venables 1995). Still, it remains to be seen what part of international income differences can be attributed to development traps (for steps in this direction, see Fafchamps and Helms 1996; Graham and Temple 2006).

See Also

- ▶ Balanced Growth
- ▶ Development Economics
- ▶ External Economies
- ▶ Externalities
- ▶ Multiple Equilibria in Macroeconomics
- ▶ New Economic Geography
- ▶ Returns to Scale
- ▶ Supermodularity and Supermodular Games

Bibliography

Chenery, H., S. Robinson, and M. Syrquin. 1986. *Industrialization and growth: A comparative study*. New York: Oxford University Press.

Ciccone, A. 2002. Input chains and industrialization. *Review of Economic Studies* 69: 565–587.

Fafchamps, M., and B. Helms. 1996. Local demand, investment multipliers, and industrialization: theory and application to the Guatemalan highlands. *Journal of Development Economics* 49: 61–92.

Fleming, M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65: 241–256.

Graham, B.S., and J.R.W. Temple. 2006. Rich nations, poor nations: how much can multiple equilibria explain? *Journal of Economic Growth* 11: 5–41.

Hirschman, A. 1958. *The strategy of economic development*. New Haven, CT: Yale University Press.

Krugman, P. 1993. Toward a counter-counterrevolution in development theory. In *Proceedings of the World Bank annual conference on development economics*, ed. L.H. Summers and S. Shah. Washington, DC: World Bank.

Krugman, P. 1994. The fall and rise of development economics. In *Rethinking the development experience: Essays provoked by the work of Albert O. Hirschman*, ed. L. Rodwin and D.A. Schon. Washington, DC: Brookings Institution.

Krugman, P., and A.J. Venables. 1995. Globalization and the inequality of nations. *Quarterly Journal of Economics* 110: 857–880.

Murphy, K.M., A. Shleifer, and R.W. 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.

Okuno-Fujiwara, M. 1988. Interdependence of industries, coordination failure and strategic promotion of an industry. *Journal of International Economics* 25: 25–43.

Rodriguez-Clare, A. 1996. The division of labor and economic development. *Journal of Development Economics* 49: 3–32.

Rosenstein-Rodan, P. 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: 143–151.

Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.

Lintner, John Virgil (1916–1983)

J. Fred Weston

Keywords

Asset pricing; Capital asset pricing model; Capital formation under inflation; Dividend policy; Financial markets theory; Lintner, J. V.; Merger analysis

JEL Classifications

B31

Lintner was born in Lone Elm, Kansas. He received the Ph.D. at Harvard University in 1946, becoming a member of the faculty a year earlier. He remained a member of the Harvard faculty throughout his career and was designated the George Gund Professor of Economics and Business Administration in 1964, with a joint appointment in the Business School and the Faculty of Arts and Sciences in Economics.

The contributions by John Lintner that are most frequently cited in the economic literature

involve asset pricing, dividend policy, mergers, and capital formation under inflation. Along with others, Lintner was one of the independent creators of the modern theory of asset pricing. This model is usually referred to as the capital asset pricing model (CAPM) which holds that the equilibrium rates of return on all risky assets are a function of their covariance with the returns on the market portfolio.

In addition to his major contribution to the creation of the modern theory of financial markets, Lintner wrote the seminal articles on dividend policy which provided the foundations for further research and remain the basic references on the subject.

Mergers represented the third area of important contributions. An early study focused on the impact of taxes on mergers (Butters et al. 1951). One important impact of taxes documented was the sale of companies to convert an earnings stream that would otherwise be subject to personal income tax rates to capital gains which would be taxed at lower rates. His later studies of mergers developed an analysis of the historical influences on mergers during the major merger movements of the United States. In addition, a theoretical rationale for pure conglomerate mergers was also developed (1971).

Many aspects of Lintner's interest in capital formation under inflation were brought together in his Presidential Address to the American Finance Association in December 1974 (1975). His subsequent work sought to develop further the major themes which he had set forth in his Presidential Address.

Selected Works

1951. (With J.K. Butters and W.L. Cary.) *Effects of taxation on corporate mergers*. Boston: Harvard Business School.

1956. Distribution of incomes of corporations among dividends, retained earnings and taxes. *American Economic Review* 46: 97–113.

1964. Optimal dividends and corporate growth under uncertainty. *Quarterly Journal of Economics* 78: 49–95.

1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

1965. Security prices, risk, and maximal gains from diversification. *Journal of Finance* 20: 587–615.

1969. The aggregation of investors' diverse judgments and preferences in purely competitive security markets. *Journal of Financial and Quantitative Analysis* 4: 347–400.

1971. Expectations, mergers, and equilibrium in purely competitive securities markets. *American Economic Review* 61(2): 101–111.

1975. Presidential address. American Finance Association Annual Meeting, San Francisco, California, 29 December 1974. *Journal of Finance* 30: 259–280.

Liquidity

A. B. Cramp

Liquidity is a highly complex phenomenon. Its concrete manifestation is powerfully affected by changes in financial institutions and practices, which have been occurring with extraordinary rapidity in recent decades. It calls for analysis both at the microeconomic and the macroeconomic level, with unusually strong dangers of committing fallacies of composition. It needs to be conceptualized both *ex ante* and *ex post,* involving recognition that the latter perspective alone facilitates statistical estimation, while the former is more relevant to transactors' wealth-holding and expenditure decisions. Together, these factors render extremely difficult a definitive answer to the major policy-related issue, namely the extent to which liquidity weakens the Quantity Theory link between 'money' stocks and expenditure flows.

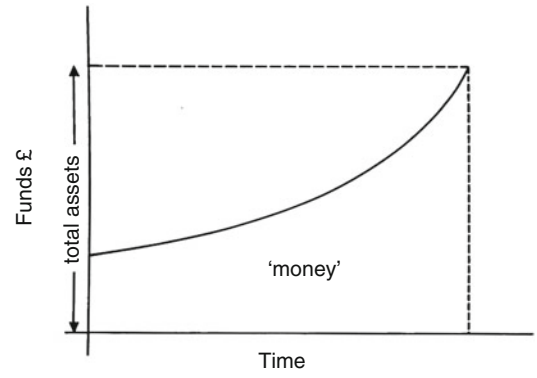
As this statement of the issue implies, debate focuses initially mainly at the macroeconomic level, and mainly on financial (as opposed to

'real') assets. These have been classified by Hicks (1967) into the categories of (1) *running assets,* required by transactors for the maintenance of their activity; (2) *reserve assets,* held to facilitate flexibility of response to ill-foreseen change in economic stimuli; (3) *investment assets,* held for their yield.

Category (1) includes 'money' balances needed to satisfy Keynes's transactions motive to liquidity. But in addition to these claims on (primarily?) banks, it also includes claims on non-financial entities in the form of trade credit, representing goods which have been sold but not yet paid for. Category (2) includes money balances held to satisfy Keynes's precautionary motive to liquidity, along with a familiar spectrum of liquid assets which are mostly shortterm claims on the public sector (e.g. Treasury Bills) or on non-bank financial institutions (e.g. building society deposits). Category (3) includes Keynes's speculative money balances, as well as the whole gamut of long-term claims in the form of bonds and so on.

This classification of financial assets, though heroically simplified, is adequate to facilitate discussion and assessment of the three main conceptualizations of liquidity which have emerged in the course of efforts at clarification (see Newlyn 1962). The first of these has been labelled *maturity.* Treating 'money' as an asset having zero life to maturity, and on the (strong) simplifying assumption that all assets possess specific maturity dates, one may notionally construct a 'maturity curve' showing the cumulative total of assets due to mature by various future dates (Fig. 1). For a given asset total, the higher is the intercept of this curve, and the shallower the gradient, the more liquid is the economy's position – because the closer assets are to maturity, the greater in general is the possibility of realizing them before maturity without risk of significant capital loss. It would follow that the more liquid an economy is in this sense, the greater is its capacity to sustain varying output levels without inhibition from interest-rate volatility and associated changes in the market value of a given asset stock.

Such an account presumes that 'money' plays no unique role in the process of acquisition and disposal of financial assets. But in reality, of course, non-money assets are not normally realized, and used to finance spending, without first being exchanged for money balances. This pivotal intermediary role of money is recognized by the second of the three major liquidity concepts, namely *easiness*: this has been defined as the ratio of the stock of money balances (not to the stock of wealth but) to the flow of output, that is, *M/Y.* The apparent implication is that a high ratio would facilitate expansion of output if adequate incentive existed, while a low ratio would tend to inhibit expansion and possibly enforce contraction. Such an implication is consonant, of course, with the Quantity Theory tradition. In assessing its practical validity, it is necessary to indicate doubts arising from a 'Liquidity Theory' perspective of the kind adumbrated most powerfully, perhaps, by Gurley and Shaw (1960).



Liquidity, Fig. 1

These doubts are of three main kinds. First, it has proved impossible to define money in a manner that commands universal (or even widespread) assent, and enables it to be distinguished clearly from what have been variously labelled liquid assets, near-moneys, or money substitutes (see Sayers 1960). This is true of the situation in (financially sophisticated) economies at any particular time, and the difficulty is compounded when attention is directed to changes in institutional structures and practices, changes always occurring, more rapidly or less. Historically, bank notes and bank deposits were initially regarded as means of economizing on holdings of balances of 'real money', or metallic coin.

First bank notes, then demand deposits, were admitted to the money category. But what of bank time deposits, holders of which could normally suppose that banks would honour their cheques, in effect treating the deposits as belonging to the demand category, usually without substantial penalty? And if bank time deposits be regarded as money, how do they differ fundamentally from, say, building society deposits normally held on similar terms and for similar purposes? And if money is so ineradicably slippery conceptually, can it be so important an entity, in developed economies at least, as Quantity Theory reasoning suggests?

These considerations are closely related to the second kind of doubt, which concentrates on the notion, central to modern Quantity Theory reasoning, of a firm and identifiable demand for money, functionally related to a relatively small number of identifiable variables (e.g. wealth stocks, asset yields). If monetary assets are held in each of Hicks's three categories already mentioned, and within each category are grouped with alternative assets which may be more or less closely substitutable, there would seem in principle to be considerable scope for portfolio adjustment by transactors, to offset any potential effects of monetary stringency on spending plans.

The force of these two kinds of doubt might be weakened, were it true that the supply of particular classes of asset, to which the label 'money' might be affixed, proved to be unresponsive to changing private sector demand, or in other words was determined 'exogenously'. It might then follow that this supply, particularly if its 'givenness' were reinforced by restrictive monetary policy measures, would act as a significant brake on the possibilities of portfolio reshuffling mentioned in the previous paragraph, because a situation might be reached in which wealth-holders were unable to switch into 'money' assets on nonpenalty terms, or even at all. In fact, however, our third kind of doubt centres precisely on the claim that the supply of *all* assets, including those which might be called 'money', is essentially subject to endogenous rather than exogenous determination. Before investigating this claim, however, it will be well to introduce our third major liquidity concept, known as *financial strength.*

The explication of this concept calls for recognition of two further complications for financial analysis, largely avoided so far in this article. The first of these is the distinction between the public and the private sectors of the economy. The second is the recognition, perhaps rather belated, that financial claims which represent assets to their holders also represent liabilities to their issuers. With these points in mind, we may approach a simple analysis of the financial strength of, initially, an individual private sector 'transactor' – whether person/family, company/organization, or other entity.

Beginning on the asset side of such a transactor's balance sheet, we may regard holdings of claims on the government (g), and on other private sector entities (a_p), both measured at market rather than nominal values, as unambiguously contributing to the transactor's financial strength (Z) – which can thus be represented by $g + a_p$.

However, it is necessary to make some offset on account of the transactor's liabilities, presumed for simplicity to be entirely due to private sector bodies, and which we may label l_p. So we have $Z = g + a_p - l_p$. But the offset arguably need not include *all* such liabilities, for the transactor may be regarded as being content to incur some volume of liabilities, as having a 'propensity to owe', ω. This propensity, however, must be limited by reference *(inter alia)* to the size of the (present and prospective) income streams from which debt may be serviced; it may thus be expressed as a proportion of income, ωY. So the final expression for Z is $g + a_p - (l_p - \omega Y)$.

This, to repeat, is the expression for the individual transactor. Aggregating for the whole economy, on the (debatable) assumption that asset-holders' and liability-issuers' reactions to growth of claims are equal and opposite, we arrive by cancellation at an expression for the financial strength of the private sector as $g - \omega Y$. (It will be noted that this approach treats g as being, in Gurley and Shaw's terminology, 'outside money', on the assumption – again debatable but probably often roughly valid – that *government* spending is not inhibited by the size of its existing liabilities.)

But just one more layer of complexity is unavoidable if we are to achieve even provisional

approximation to an extraordinarily confusing reality. We must recognize that much, perhaps the bulk, of private-sector debt will be owed to financial institutions, so that the picture is seriously incomplete unless we incorporate some notion, however simplified, of the conditions on which financial institutions will lend, and in particular of the elasticity of supply of credit – in response to changes in demand for credit, and in interest rates.

It is the contention of many theorists that financial intermediaries typically give priority to meeting the demands of their private sector customers, absorbing volatility in credit demand by (a) attracting new deposits at interest rates which rise only gently because of a high elasticity of substitution among reserve assets; and (b) permitting their reserves, largely in the form of holdings of g, to fluctuate. The result is that credit supply is seen as being highly elastic to private sector demand. Moreover, according to the so-called 'new view' of banking theory, on which see Tobin (1963), point (b) at least is true of banks as well as of other financial institutions. The conclusion is that the supply of bank deposits, the stock residue of previous bank credit flows, is also essentially determined 'endogenously' rather than 'exogenously'. The implication is that the supply of money does *not* automatically act as a significant brake on possibilities of portfolio reshuffling indicated above, and that monetary policy operating through market methods as opposed to direct controls would be hard-pressed to change the situation significantly. This implication would, however, be liable to break down in the case of a liquidity crisis following a strong boom; but such conditions in any case enforce relaxation of tight money policies.

Putting together the rather complex considerations we have outlined, the case for believing that liquidity in modern economies does weaken the Quantity Theory is arguably very strong. In the face of monetary stringency, transactors can sustain spending flows by reshuffling asset portfolios. Some part of this reshuffling will provide financial intermediaries with increased lending powers. Banks tend to maintain private sector lending flows by lending less to the public sector. In Quantity Theory language, money is in quite elastic supply, the velocity of circulation is volatile enough to offset such monetary restriction as the authorities may achieve, and the much-derided argument of the Radcliffe Report (1959) concerning the liquidity-related weakness of reasonably gentle monetary policy using market methods is essentially correct.

See Also

- ▶ Capital, Credit and Money Markets
- ▶ Central Banking
- ▶ Finance
- ▶ Financial Intermediaries

Bibliography

Gurley, J.G., and E.S. Shaw. 1960. *Money in a theory of finance*. Washington, DC: Brookings Institution.

Hicks, J.R. 1967. *Critical essays in monetary theory*. Oxford: Oxford University Press.

Newlyn, W.T. 1962. *The theory of money*. Oxford: Oxford University Press.

Report of the Committee on the Working of the Monetary System. 1959. (The Radcliffe report) London: HMSO, Cmnd. 827.

Sayers, R.S. 1960. Monetary thought and monetary policy in England. *Economic Journal* 70, December: 710–724.

Tobin, J. 1963. Commercial banks as creators of 'money'. In *Banking and monetary studies*, ed. D. Carson. Homewood: Richard D. Irwin.

Liquidity Constraints

Stephen D. Williamson

Abstract

Liquidity constraints affect the ability of an economic agent to exchange his or her existing wealth for goods and services or for other assets. These constraints arise because of frictions, including private information, limited commitment, transactions costs, and spatial considerations.

Keywords

Cash-in-advance model; Consumption smoothing; Contingent claims markets; Endowments; Expected utility; Incomplete markets; Law of large numbers; Limited commitment; Liquidity constraints; Precautionary savings; Private information; Risk sharing; Search and matching models of monetary exchange

JEL Classifications

D4; D10

A Benchmark Model

To explain what liquidity constraints are, and their implications for economic activity, it is useful to start with a simple benchmark model. Suppose a world with a continuum of households having unit mass. Time is indexed by $t = 0, 1, 2, \ldots$, and household i has preferences given by

$$E_0 \sum_{t=0}^{\infty} \beta^t u(c_{it}),$$

where E_0 is the expectation operator conditional on period 0 information, $0 < \beta < 1$, c_{it} is consumption, and $u(\cdot)$ is twice continuously differentiable, strictly concave, and has the property that $u'(0) = \infty$. Each household receives a random endowment of the perishable consumption good at the beginning of each period. That is, household i receives an endowment y_{it} in period t where y_{it} is assumed to be independent and identically distributed across households and over time. Assume that $y < \underline{y} < \overline{y}$, where $0 < \underline{y} < \overline{y}$. The law of large numbers then implies that the aggregate endowment is a constant, which we will denote by y. Therefore, this is an economy with no aggregate risk, but each household faces idiosyncratic risk associated with its endowment shocks.

Now, suppose that this economy has a complete set of markets. One market structure that gives completeness is contingent claims markets that open at $t = 0$ before households receive their period 0 endowments. All households trade on these markets, and a particular contingent claims market involves trade in claims to the consumption good deliverable at a particular date only under a particular realization for the path of endowment shocks for all households up to that date. Given this complete set of markets, what will be the equilibrium allocation of consumption across households at each date? All households are identical at the first date, and the result will be that, in equilibrium, $c_{it} = y$ for all i and t. The complete set of contingent claims markets provides perfect insurance for households. That is, they are able to share their risk efficiently, in that each household can shed the idiosyncratic risk associated with its endowment shocks. Indeed, the resulting equilibrium allocation of consumption is Pareto optimal.

Models with complete markets have proved to be very useful in economics, for example in the theory of asset pricing and in business cycle modelling. However, there are many applications where it is necessary that we depart from the complete markets paradigm, and the liquidity constraints literature is one such set of applications. To think about liquidity constraints we need to seriously address the frictions that will cause markets to work differently than in the complete markets case, and in some instances will cause some markets to shut down altogether. In the following sections we will explore some key departures from our benchmark model that illustrate the role of liquidity constraints.

Incomplete Markets: A Bewley Model

One approach to studying market incompleteness is to simply eliminate markets in the model under consideration, without asking questions about the underlying frictions which would cause incomplete markets. Bewley (1977) was a pioneer in this area, and Aiyagari (1994) provides a particularly clear treatment of the implications of incomplete markets.

As an example of the Bewley approach, suppose in our benchmark model that there is only

one asset market, a market for non-contingent bonds on which trading occurs each period. Households can borrow and lend on this bond market. Assume that each bond is a one-period financial instrument. In period t, a bond sells for one unit of consumption goods, and is a promise to pay $1 + r_{t+1}$ units of consumption goods in period $t + 1$. Since there is no aggregate risk, there will exist a steady state competitive equilibrium where $r_{t+1} = r$, a constant, for all t.

We now need to write down the series of constraints that a household faces in the steady state equilibrium. The first of these is the sequence of budget constraints

$$c_{it} + b_{i,t+1} = y_{it} + (1 + r)b_{i,t},$$

for $t = 0,1,2,\ldots$, where b_{it} is the quantity of bonds acquired by household i in period t, and $b_{i0} = 0$ for all i. Typically in models of this type, there is also a borrowing constraint added, which could take the form

$$b_{it} \geq \underline{b}. \tag{1}$$

Constraint (1) serves a technical purpose, in that it prevents a household from borrowing an infinite amount so as to finance infinite consumption. Further, the constraint will affect the household's ability to smooth consumption over time in the face of fluctuating income. Constraint (1) is a kind of liquidity constraint, as it potentially prevents the household from borrowing against its lifetime wealth.

A competitive equilibrium will have the property that the bond market clears, that is the net stock of bonds in the population is zero in each period. This model is a special case of Aiyagari (1994), and so his results apply here. With Aiyagari's regularity conditions on $u(\cdot)$, a steady state competitive equilibrium will have the property that $r < \frac{1}{\beta} - 1$, that is, the equilibrium real interest rate is less than the rate of time preference. This reflects a precautionary savings motive, in that households wish to hold bonds to self-insure against having a string of bad luck, which in this case would be a string of low endowment shocks. Over time, a household will tend to increase its stock of bonds when its endowment is large, and to decrease the stock of bonds when its endowment is small. What we will observe in equilibrium is some distribution of bonds and consumption across the population of households. Households who have had good luck will tend to have a larger stock of bonds and higher consumption than those households who have had bad luck. The competitive equilibrium is therefore not in general Pareto optimal.

Another related application, from Bewley (1980), is to suppose that the single asset that is traded is money. For example, suppose that there is a fixed stock of money, M, for all t. Let P_t denote the price level in period t, and consider the steady state equilibrium where $P_t = P$, a constant, for all t. For the household, we can just reinterpret its constraints, in that $b_{i,t+1}$ is the real quantity of money carried over by the household into period $t + 1$, and $\underline{b} = 0$ as the household's money balances cannot fall below zero. An individual household in this set-up is even more severely liquidity-constrained than was the household in the Bewley model with borrowing and lending above. This is because the household cannot borrow at all, and cannot hold interest-bearing assets. Note that, in this monetary model, a household need only use money to buy consumption goods if it wishes to consume more than its endowment. Money is essentially held for insurance purposes, so as to smooth consumption over time.

Cash-In-Advance

The idea for the basic cash-in-advance model seems to come from Clower (1967), but the important initial modelling work was done mainly by Robert Lucas, with a key contribution being Lucas (1980). Most cash-in-advance applications begin with the view that the basic frictions that might give rise to cash-in-advance constrained households need not be modelled, and that it is useful to proceed from the premise that money is necessary to purchase some goods and services.

Here, suppose in our basic model that there are no assets other than money, and that the only

exchanges are trades of money for goods. Assume that a household's purchases of goods during the current period must be financed with money carried over from the previous period, and also suppose that the household cannot consume its own endowment. Let m_{it} denote the nominal money balances that household i has at the beginning of period t, and let P_t denote the price level. Then, the household's budget constraint in period t is

$$P_t c_{it} + m_{i,t+1} = P_t y_{it} + m_{it}. \tag{2}$$

The cash-in-advance constraint for the household is

$$P_t c_{it} \leq m_{it}. \tag{3}$$

Thus, constraint (3) is another type of liquidity constraint. In this case, the interpretation is that some class of assets, which we refer to here as money, is necessary to carry out goods market spot exchanges.

Now, suppose that there is a fixed nominal stock of money M. Also, suppose that in equilibrium constraint (3) binds for each household i. Then, since in equilibrium the entire stock of money is held by households at the beginning of period and is spent to purchase the aggregate endowment, y, the equilibrium price level is

$$P_t = \frac{M}{y}$$

for all t. Then, given (2) and (3) with equality, we have

$$m_{i,t+1} = M\frac{y_{it}}{y},$$

which then implies, from (3) with equality, that

$$c_{it} = y_{i,t-1}.$$

Therefore, in this environment, households have essentially no ability to smooth consumption relative to income, as a result of this extreme type of liquidity constraint. The distribution of consumption across households in period t is determined by the distribution of income across households in the previous period.

Economists who are serious about monetary theory often treat cash-in-advance models with some disdain (see, for example, Wallace 1996). As they see it, the problem is not that one cannot write down a model that is explicit about frictions and gives rise to cash-in-advance as an endogenous phenomenon. For example, suppose that we modify our benchmark model to permit an absence-of-double-coincidence friction of the type considered by early monetary theorists such as Jevons (1875). That is, assume that households are of N types, with measure $\frac{1}{N}$ households of each type, where type is indexed by $j = 1, 2, \ldots, N$. Type j households are endowed with good j, and consume the good which is endowed to type $j + 1$, modulo N. Further, suppose that a household has two members, a shopper that takes money from the household to buy goods in another market each period, and a seller who stays at home to sell the household's endowment. There are N distinct markets, and in a given period a shopper from a household of type j goes to market $j + 1$, modulo N, with money to buy goods, while a the seller stays behind and sells goods in market j. Note that this is still not enough to give us cash-in-advance, as we need to close off the possibility of credit arrangements among households which could take place through centralized communication, as is made clear in Kocherlakota (1998). Credit can be shut down by assuming that no communication is possible across markets, with buyers and sellers in a given market having no information about each other, beside the fact that sellers have identifiable goods and buyers have identifiable money balances. With competitive pricing in each of the N markets, we get exactly the set-up outlined above in this section, with a cash-in-advance constraint for each household. Given symmetry, there is an equilibrium where prices are the same in every market, and so the equilibrium allocation of consumption is identical to what was specified above.

The key problem that must be addressed in cash-in-advance environments involves what happens when there are other assets than money. For example, if we permit borrowing and lending by

households, why is it that goods cannot be purchased with credit? How can money be dominated in rate of return by other assets? Why is it that government bonds, for example, are not used in transactions rather than money? Many cash-in-advance applications leave these questions unanswered.

Random Matching

A useful way to extend our benchmark model at this point is to expand on the explicit cash-in-advance environment above to relate it more directly to the literature on monetary search and matching. The seminal work in this literature is by Jones (1976) and Kiyotaki and Wright (1989).

Suppose as above that there is a double coincidence problem, but here assume that there is one agent in a household, and that each household is randomly matched with one other household each period. Households produce different goods, and no household can consume its own endowment. Now, for a given household, assume that the probability is a that it is matched with another household whose goods it consumes, with the other household not wanting its goods (a single coincidence meeting). As well, assume that there is a probability γ that a household is matched with another household and there is a double coincidence of wants – each household consumes the other's goods. Suppose that $\alpha > 0$, $\gamma > 0$, and $2\alpha + \gamma < 1$. Suppose that a household in a bilateral match has no information about the other household, except that it can observe its quantity of money balances and its endowment. Thus, exchange can only involve bilateral exchanges of goods and money.

Now, suppose that household i and household k are matched. There is probability a that household i is a seller and k is a buyer. In this case, we have $c_{it} = 0$, $c_{kt} = y_{it}$, and

$$m_{i,t+1} = m_{it} + m(y_{it}, m_{it}, m_{kt}),$$
$$m_{k,t+1} = m_{k,t} - m(y_{it}, m_{it}, m_{kt}),$$

where $m(y_{it}, m_{it}, m_{kt})$ is the quantity of money exchanged for the y_{it} units of goods given up by the seller when the seller has m_{it} units of money and the buyer has $m_{kt.}$ units of money. As money balances must be non-negative, we have

$$-m_{it} \leq m(y_{it}, m_{it}, m_{kt}) \leq m_{kt} \tag{4}$$

and these constraints are essentially liquidity constraints. Similarly, with probability α, household i is the buyer and k is the seller, in which case $c_{it} = y_{kt}$, $c_{kt} = 0$, and

$$m_{i,t+1} = m_{it} - m(y_{kt}, m_{kt}, m_{it}),$$
$$m_{k,t+1} = m_{k,t} + m(y_{kt}, m_{kt}, m_{it}),$$

with

$$-m_{kt} \leq m\ (y_{kt}, m_{kt}, m_{it}) \leq m_{it}. \tag{5}$$

Finally, with probability γ there is a double coincidence, and household i and k exchange goods, so that $c_{it} = y_{kt}$, $c_{kt} = y_{it}$, and

$$m_{i,t+1} = m_{it} + b(y_{it}, y_{kt}, m_{it}, m_{kt}),$$
$$m_{k,t+1} = m_{k,t} - b(y_{it}, y_{kt}, m_{it}, m_{kt}),$$

where

$$-m_{it} \leq b\ (y_{kt}, y_{kt}, m_{it}, m_{kt}) \leq m_{kt}. \tag{6}$$

Here, $b(y_{it}, y_{kt}, m_{it}, m_{kt})$ is the quantity of money passed from household k to household i, which depends on the money balances and endowments of each household.

Note that this environment will give a clear sense in which money improves the equilibrium allocation. If money is not valued, then households can trade only when there is a double coincidence of wants, and this could severely limit exchange possibilities. In principle, the constraints (4, 5 and 6) will matter for the equilibrium allocation in important ways. However, the model as we have laid it out is quite intractable. It is possible to use a bargaining approach, as for example in Trejos and Wright (1995) or Shi (1995), to determine how much money is transferred in each type of match, but the key problem is in tracking the distribution of money balances in the population over time.

In some of the monetary search and matching literature, tractability is achieved through assuming that money and goods are indivisible (Kiyotaki and Wright 1989) or that money is indivisible and goods are divisible (Trejos and Wright 1995; Shi 1995), and that there is an inventory constraint on money holdings. If a household can hold only one unit of money or nothing, and money is never disposed of, then the quantity of money outstanding tells us how many households have it and how much, and how many do not have it. Models with indivisible money yield some insights, but they are extremely awkward for dealing with some types of policy questions, such as those involving money growth and the effects of inflation. Some recent progress in the development of tractable search models of divisible money was achieved by Lagos and Wright (2005), who use a quasilinear utility setup with labour supply and alternating periods of centralized meeting and search. This type of model yields a result where, in the periods when centralized meeting takes place, economic agents optimally redistribute money among themselves in such a way that the distribution of money balances becomes degenerate. Recent research using this type of model (for example, Williamson 2006; Berentsen et al. 2005) has been quite productive.

Private Information and Limited Commitment

As an alternative to shutting down markets in an ad hoc fashion, imposing borrowing constraints, assuming cash-in-advance constraints, or making extreme informational assumptions that shut down all trade except monetary exchange, there are available approaches to facing frictions head-on that lead to incomplete insurance and imperfect credit. These approaches involve economies with private information and limited commitment.

A well-developed approach to dealing with private information frictions in large economies follows the pioneering work of Green (1987), Atkeson and Lucas (1993) and others. Extending our benchmark model, suppose now that endowments are private information. In our baseline environment, we know that if endowments are public information, then a Pareto optimal allocation that treats households identically has $c_{it} = y$ for all i,t. What is optimal from a social planner's point of view under private information?

It is clear that private information implies that the $c_{it} = y$ allocation cannot be implemented by the social planner. To see this, note that to achieve this allocation requires that household i make a transfer of $y_{it} - y$ to the planner in period t. But it would then be incentive compatible for every household in every period to report that its endowment was $\underline{y}$, and so the planner could not achieve this allocation.

Following Green (1987) and Atkeson and Lucas (1995), one can solve for an optimal private information allocation by recursive methods. The state variable for any household is w_{it}, which is the level of expected utility promised to the household as of the beginning of period t. At the beginning of period t, the household reports its endowment y_{it} to the social planner, and it must be optimal for the household to report the truth (that is, the allocation must be incentive compatible). The planner delivers consumption $c(w_{it}, y_{it})$ to the household, which depends on its state and reported endowment, and promises expected utility $w(w_{it}, y_{it})$ for next period. There is a functional equation that solves for a cost function $V(w_{it})$, which is the cost to the social planner of delivering expected utility w_{it} to a particular household. On the right-hand side of this functional equation is a cost minimization problem, and the minimization is subject, first, to a promise-keeping constraint, which is

$$w_{it} = \int \{u[c(w_{it}, y_{it})] + \beta w(w_{it}, y_{it})\} dF(y_{it}),$$

where $F(y_{it})$ is the distribution function for y_{it}. The remaining constraints are incentive compatibility constraints, written as

$$u[c(w_{it}, y_{it})] + \beta w(w_{it}, y_{it}) \geq u\left[c\left(w_{it}, \tilde{y}\right) + y_{it} - \tilde{y}\right] + \beta w(w_{it}, \tilde{y}),$$

for all $y_{it}, \tilde{y} \in \underline{y}, \overline{y}]$. The optimal allocation will typically have the property that some incentive

compatibility constraints bind. For efficient risk sharing, we want households with high (low) endowments to be making positive (negative) transfers to the social planner. To accomplish this in an incentive compatible manner requires that households with high (low) endowments receive increases (decreases) in their future expected utility promises. Thus, the distribution of consumption will tend to fan out over time. Under some conditions, a vanishing fraction of households will ultimately consume the entire endowment. However, under other conditions there will be a limiting distribution of expected utility promises with mobility and a lower bound on expected utilities. If a household hits this lower bound (which is not absorbing), then this is much like having a borrowing constraint bind for this household. Thus, this type of set-up can yield what are essentially endogenous borrowing constraints or liquidity constraints.

An alternative approach to modeling frictions in a serious way is to assume some form of limited commitment. One approach to limited commitment is that of Kehoe and Levine (1993), which has elements of competitive equilibrium. Extending our benchmark model to illustrate the flavour of this modelling approach, suppose that there is only one type of intertemporal trade, involving one-period bonds, and that we wish to study a steady state where the real interest rate is a constant, r. Suppose that the key friction here is that a household may decide strategically to repudiate its debt, in which case it would be barred from the credit market for ever and would then consume its own endowment for ever. Thus, if a household does not repudiate its debt, then its budget constraint is given by

$$c_{it} + b_{i,t+1} = y_{it} + (1 + r)b_{it}, \tag{7}$$

where $b_{i,t+1}$ is the quantity of one-period bonds acquired in period t that each pay off $1 + r$ units of consumption in period $t + 1$. Let $v(b_{it}, y_{it})$ denote the expected utility of the household at the beginning of period t as a function of the household's asset position and endowment, determined by the functional equation

$$v(b_{it}, y_{it}) = \max_{c_{it}, b_{i,t+1}} \left[u(c_{it}) + \beta \int v(b_{i,t+1}, y_{i,t+1}) dF(y_{i,t+1}) \right]$$

subject to (7). To insure that the household does not repudiate its debt in equilibrium requires that the value of not repudiating is no smaller than the value of repudiating, or

$$v(b_{it}, y_{it}) \geq u(y_{it}) + \frac{\beta}{1 - \beta} \int u(\hat{y}) dF(\hat{y}). \tag{8}$$

Note that constraint (8) is another type of borrowing constraint or liquidity constraint. Typically, $v(b_{it}, y_{it})$ must be strictly increasing in b_{it} and so, given y_{it}, there will be some critical value of b_{it} for which the constraint binds. Thus, lenders cannot lend too much to a particular household, as doing so would imply debt repudiation.

Kocherlakota (1996) takes a somewhat different approach by examining a two-agent problem with limited commitment. In his set-up, two infinite-lived agents work out a risk-sharing arrangement subject to limited commitment. Kocherlakota's problem does not have some of the loose ends found in Kehoe and Levine (1993). In the Kehoe and Levine model, we are forced to accept an incomplete markets view of the world with no explanation for why the markets are missing, and it is not clear how credit market participants coordinate to discipline agents who repudiate their debts.

Aiyagari and Williamson (2000) integrate private information and limited commitment with a Bewley model of monetary exchange to study the relationship between money and credit. Credit arrangements are constrained by private information considerations, and if agents defect from credit arrangements their alternative is to be liquidity constrained in the manner of a Bewley-type consumer.

See Also

- ▶ Aiyagari, S. Rao (1952–1997)
- ▶ Incomplete Markets
- ▶ Lucas, Robert (Born 1937)
- ▶ Money
- ▶ Money and General Equilibrium

Bibliography

Aiyagari, R. 1994. Uninsured idiosyncratic risk and aggregate saving. *Quarterly Journal of Economics* 109: 659–684.

Aiyagari, R., and S. Williamson. 2000. Money and dynamic credit arrangements with private information. *Journal of Economic Theory* 91: 248–279.

Atkeson, A., and R. Lucas. 1993. On efficient distribution with private information. *Review of Economic Studies* 59: 427–453.

Atkeson, A., and R. Lucas. 1995. Efficiency and inequality in a simple model of unemployment insurance. *Journal of Economic Theory* 66: 64–88.

Berentsen, A., G. Camera, and C. Waller. 2005. The distribution of money balances and the nonneutrality of money. *International Economic Review* 46: 465–488.

Bewley, T. 1977. The permanent income hypothesis: A theoretical formulation. *Journal of Economic Theory* 16: 252–292.

Bewley, T. 1980. The optimum quantity of money. In *Models of monetary economies*, ed. J. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Clower, R. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6: 1–8.

Green, E. 1987. Lending and the smoothing of uninsurable income. In *Contractual arrangements for intertemporal trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Jevons, W.S. 1875. *Money and the mechanism of exchange*. London: Appleton.

Jones, R. 1976. The origin and development of media of exchange. *Journal of Political Economy* 84: 757–775.

Kehoe, T., and D. Levine. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60: 865–888.

Kiyotaki, N., and R. Wright. 1989. On money as a medium of exchange. *Journal of Political Economy* 97: 927–954.

Kocherlakota, N. 1996. Implications of efficient risk-sharing without commitment. *Review of Economic Studies* 63: 595–610.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81: 232–251.

Lagos, R., and R. Wright. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113: 463–484.

Lucas, R. 1980. Equilibrium in a pure currency economy. In *Models of monetary economies*, ed. J. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Shi, S. 1995. Money and prices: A model of search and bargaining. *Journal of Economic Theory* 67: 467–496.

Trejos, A., and R. Wright. 1995. Search, bargaining, money, and prices. *Journal of Political Economy* 103: 118–141.

Wallace, N. 1996. A dictum for monetary theory. In *Foundations of research in economics: How do economists do economics?* ed. S.G. Medema and W.J. Samuels. Cheltenham: Edward Elgar.

Williamson, S. 2006. Search, limited participation, and monetary policy. *International Economic Review* 47: 107–128.

Liquidity Effects, Models of

Chris Edmond and Pierre-Olivier Weill

Abstract

An exogenous increase in the money supply is typically followed by a temporary fall in nominal interest rates. Flexible price macroeconomic models argue that this *liquidity effect* arises because asset markets are segmented. That is, only a fraction of the agents are present in the bond market when the central bank conducts an open market operation. However, to be quantitatively successful, segmented markets models assume frictions that are too large to be interpreted literally in terms of constraints faced by real-world firms and households. An important open question is: can a complicated array of microeconomic frictions imply one large aggregate friction of this kind?

Keywords

Asset market frictions; Asset market segmentation; Cash-in-advance models; Fisher effect; Inflation; Inflation expectations; Liquidity effects; Long-Horizon interest rates; Monetary transmission mechanism; Money supply; Nominal interest rates; Open-Market operations; Real business cycle; Real interest rates; Short-Horizon liquidity effects; Velocity of circulation

JEL Classifications

D4; D10

In macroeconomics, the term *liquidity effect* refers to a fall in nominal interest rates following an

exogenous persistent increase in narrow measures of the money supply. According to the classical *Fisher effect*, however, an exogenous persistent increase in money is predicted to increase expected inflation and so increase nominal interest rates. Friedman (1968) argues that, in practice, both forces operate: a persistent increase in the money supply both reduces nominal interest rates and increases expected inflation so that the real rate – nominal minus expected inflation – also falls. Friedman (1968, pp. 5–7) speculates that nominal and real rates may fall below their typical levels for up to a year, but, over time, rates will then tend to increase before tending to the levels consistent with the inflation generated by the original monetary impulse.

Empirical macroeconomists have interpreted Friedman (1968) as follows. At long horizons real interest rates are determined by 'fundamentals' including the rate at which households discount the future and average productivity growth. Consequently, we should expect that long-horizon real interest rates are relatively stable and are unaffected by transitory monetary disturbances. Long-horizon nominal interest rates are this stable real rate plus expected inflation. At short horizons, however, Friedman's (1968) argument suggests that real and nominal interest rates are both volatile and positively correlated. His argument also suggests that short-horizon real rates and expected inflation are negatively correlated (Barr and Campbell 1997, provide evidence consistent with this interpretation and Cochrane 1989, provides specific evidence for liquidity effects at short horizons).

Perhaps the easiest way to interpret Friedman (1968) is in terms of the following market equilibrium scenario. Suppose that a monetary authority increases the money supply by conducting an unexpected outright purchase of bonds (an *open market operation*). At short horizons, nominal interest rates fall so that households are willing to hold a smaller quantity of bonds and a larger quantity of money. But this is only a partial equilibrium effect. As households spend their increased money holdings on goods, the price level increases and so real balances do not rise as fast as nominal balances. This general equilibrium effect mitigates the need for the nominal interest rate to fall. In many simple monetary models, households tend to spend money so 'fast' that the general equilibrium price level effect can completely overturn the partial equilibrium effect.

A textbook cash-in-advance (CIA) model with a constant aggregate endowment of goods ('output') and identically and independently distributed (IID) money growth shocks provides a stark example. In this model, households immediately spend an unexpected increase in money on a fixed quantity of goods. This increases the price level one-for-one with the increase in the money supply so that real balances are unchanged. In addition, because money growth is serially uncorrelated, expected inflation is constant. Taken together, constant real balances and constant expected inflation imply that the money market clears at a constant nominal interest rate. If instead monetary growth shocks are persistent then a positive shock increases expected inflation and nominal interest rates increase. In short, there is a Fisher effect but no liquidity effect. CIA models that are carefully calibrated to empirical processes for money growth and output, such as Hodrick, Kocherlakota and Lucas (1991) and Giovannini and Labadie (1991), lead to similar conclusions, as do studies of conceptually similar production economies, such as Cooley and Hansen (1989).

We now turn to departures from the standard CIA model in which a liquidity effect dominates at short horizons while a Fisher effect dominates at long horizons. Although models with nominal rigidities are in principle capable of generating these liquidity effects, we instead focus on flexible price models in which a liquidity effect is generated by an *asset market friction* of one form or another. Each of the models we discuss – Lucas (1990), Grossman and Weiss (1983), and Alvarez, Atkeson and Kehoe (2002) – captures, albeit in different ways, some of the spirit of Friedman's (1968) intuition.

Lucas (1990) modifies the standard CIA endowment economy with a simple timing assumption: households have to allocate cash between a goods market and an asset market *before* observing the size of an open-market operation. Once that allocation has been made, there is

a fixed quantity of cash sitting in the bond market. Now consider an unexpected purchase of bonds. Relative to the supply of bonds, there is now an unexpectedly large amount of cash available to purchase assets, so bond prices increase and the nominal interest rate falls.

Fuerst (1992) and Christiano and Eichenbaum (1995) integrate Lucas's (1990) timing assumption into otherwise standard real business cycle (RBC) models. The key innovation of these papers is that, in each period, firms have to borrow cash from financial intermediaries in order to pay their workers. After a positive monetary shock, the nominal interest rate decreases so that firms find it optimal to borrow the unexpected increase in money balances. This increases firms' labour demand and increases output. Thus, these models are consistent with the commonly held view that positive monetary shocks have a positive, albeit temporary, effect on output.

A limitation of models that use Lucas's (1990) timing assumption is that the liquidity effect is very transitory even when monetary shocks are persistent. Households can adjust their allocation of cash every period. Therefore, the liquidity effect is entirely driven by serially uncorrelated 'expectational errors' in cash allocation.

We now turn to Grossman and Weiss (1983) and Alvarez, Atkeson and Kehoe (2002). These are general equilibrium models inspired by Baumol (1952) and Tobin's (1956) 'inventory-theoretic' analyses of money demand. In this class of models, two key forces influence short-horizon liquidity effects. First, at any point in time, there are always some households that participate in asset markets and some households that do not. Second, because households do not acquire cash every period, they choose to spend their money holding slowly over time. The first force alone is sufficient to generate a liquidity effect; the second force provides an *amplification* mechanism.

In this setting, an open-market increase in the money supply must, in equilibrium, be held by the subset of households that are currently participating in asset markets. Therefore, even if the price level responds one-for-one with the increase in money supply, the *share* of aggregate real balances that must be held by these households increases. Hence, the nominal interest rate falls to clear the market. Also, because they hold a larger share of real balances, these households are able to increase their share of aggregate consumption and this drives down real interest rates. So, at short horizons, there is a liquidity effect.

Moreover, if households spend their money slowly over several subsequent periods then the price level does not respond one-for-one to an increase in the money supply. Instead, the price level responds slowly. This implies that aggregate real balances rise (equivalently, in a model with constant output, *velocity* falls) and this provides a second force driving down nominal interest rates. The liquidity effect is amplified.

The influential model of Grossman and Weiss (1983) is a deterministic CIA endowment economy that exhibits both effects. Households are imperfectly synchronized and only participate in asset markets every second period. They spend money on consumption goods over two periods (Rotemberg 1984, studies a production version of essentially the same environment).

Alvarez, Atkeson and Kehoe (2002) endogenize the fraction of households that participate in asset markets. They assume that households can participate if they pay a fixed cost. If a household's individual real balances are neither too high nor too low, they do not pay the cost, do not participate in asset markets, and end up consuming their individual real balances. If their real balances are high, they pay the cost and invest money in the asset market. Similarly, if their real balances are low, they pay the cost in order to purchase goods with money invested in the asset market. The equilibrium amount of participation ends up depending on the curvature of the utility function, the expected growth rate of money and on the size of the fixed cost. For example, in a high-inflation economy almost all households pay the cost to participate in asset markets. Hence, increases in the money supply raise expected inflation and nominal interest rates as in a basic CIA model. By contrast, in a low inflation economy, more households choose not to participate and the effects of incomplete participation are larger and may be big enough to cause a liquidity effect (that is, to dominate the Fisher effect at short horizons).

To simplify their analysis, however, Alvarez, Atkeson and Kehoe (2002) set up the model so that both active and inactive households spend all their money each period. No households save money to spend on consumption over multiple periods. Therefore, velocity is constant and the price level responds one-for-one with increases in the money supply. Alvarez, Atkeson and Kehoe (2002) can therefore generate a liquidity effect but without the amplification that is provided by a (transitory) fall in velocity. Alvarez, Atkeson and Edmond (2003) provide a stochastic counterpart to Grossman and Weiss (1983) where both forces are operative (but at the cost of reverting to an exogenous timing of transactions).

Limited participation models of the liquidity effect provide a number of important qualitative insights into the co-movements of money, interest, and prices (and, to a lesser extent, output). The *quantitative* insight provided by these models is, however, more debatable. To generate realistic co-movements of money, interest and prices, calibrated models of liquidity effects need 'large' asset market frictions. It is typically difficult to interpret the calibrated friction literally in terms of constraints faced by real-world firms and households (making it difficult, in the words of Manuelli and Sargent 1988, p. 524, to 'find the people'). For example, the most successful parameterizations in Alvarez, Atkeson and Edmond (2003) require the representative household to make withdrawals of money (broadly defined) from an asset market account once every 24–36 months. Alvarez, Atkeson and Edmond (2003) defend this with an appeal to the low frequency of asset market participation observed in the *cross-section* by Vissing-Jorgensen (2002). Thus, the size of the friction is defended by appealing to the likely size of the friction facing a household representative of the US economy rather than by appealing to direct evidence of the heterogeneous frictions facing individual observations of US households.

Cole and Ohanian (2002) provide another demonstration of the difficulty of interpreting such models literally. They note that the distribution of money holdings between US firms and households has been quite unstable over the post-war period. When this observation is embedded in a model of liquidity effects, it implies a corresponding instability in the effects of money shocks on output – an instability that seems to be counterfactual.

In our opinion, these limitations should not be interpreted as reasons for rejecting models of asset market segmentation. If anything, these limitations are instead reasons for rejecting an implicit aggregation hypothesis. Traditional macro models work with relatively crude frictions that are intended to summarize a complicated array of micro frictions facing individual households and firms. For example, the literature on models of liquidity effects assumes only one level of market segmentation – either between households and asset markets, or between firms and asset markets. However, asset market segmentation seems to occur at numerous levels of financial intermediation. A large body of empirical evidence shows that phenomena consistent with market segmentation arise within the financial system – a system that might best be viewed as a collection of partially integrated and relatively specialized 'local' asset markets (see, among many others, Collin-Dufresne et al. 2001).

This evidence motivates us to ask how a collection of small segmentation frictions cumulates in the aggregate, and whether they add up to a quantitatively significant macro friction. If they do, then the models of liquidity effects that we have discussed here would indeed be natural laboratories for the analysis of the monetary transmission mechanism.

In short, we conjecture that addressing segmentation at a disaggregative level is likely to provide important empirical and theoretical insights into the relationship between patterns of intermediation in financial markets and traditional macro questions – including the size and stability of liquidity effects at short horizons and the monetary policy transmission mechanism more generally.

See Also

- ▶ Finance
- ▶ Fisher, Irving (1867–1947)
- ▶ Friedman, Milton (1912–2006)

- ▶ Inflation Expectations
- ▶ Lucas, Robert (Born 1937)
- ▶ Money Supply

Bibliography

Alvarez, F., A. Atkeson, and C. Edmond. 2003. On the sluggish response of prices to money in an inventory-theoretic model of money demand. Working Paper No. 10016. Cambridge, MA: NBER.

Alvarez, F., A. Atkeson, and P.J. Kehoe. 2002. Money, interest rates and exchange rates with endogenously segmented markets. *Journal of Political Economy* 110: 73–112.

Barr, D.G., and J.Y. Campbell. 1997. Inflation, real interest rates and the bond market: A study of UK nominal and index-linked government bond prices. *Journal of Monetary Economics* 39: 361–383.

Baumol, W.J. 1952. The transactions demand for cash: An inventory-theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Christiano, L.J., and M. Eichenbaum. 1995. Liquidity effects, monetary policy, and the business cycle. *Journal of Money, Credit and Banking* 27: 1113–1136.

Cochrane, J.H. 1989. The return of the liquidity effect: A study of the short-run relation between money growth and interest rates. *Journal of Business and Economic Statistics* 7(1): 75–83.

Cole, H.L., and L.E. Ohanian. 2002. Shrinking money: The demand for money and the nonneutrality of money. *Journal of Monetary Economics* 49: 653–686.

Collin-Dufresne, P., R.S. Goldstein, and J.S. Martin. 2001. The determinants of credit spread changes. *Journal of Finance* 56: 653–686.

Cooley, T.F., and G.D. Hansen. 1989. The inflation tax in a real business cycle model. *American Economic Review* 79: 733–748.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Fuerst, T.S. 1992. Liquidity, loanable funds, and real activity. *Journal of Monetary Economics* 29: 3–24.

Giovannini, A., and P. Labadie. 1991. Asset prices and interest rates in cash-in-advance models. *Journal of Political Economy* 99: 1215–1251.

Grossman, S., and L. Weiss. 1983. A transactions-based model of the monetary transmission mechanism. *American Economic Review* 73: 871–880.

Hodrick, R.J., N. Kocherlakota, and D. Lucas. 1991. The variability of velocity in cash-in-advance models. *Journal of Political Economy* 99: 358–384.

Lucas, R.E. Jr. 1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10: 335–359.

Lucas, R.E. Jr. 1990. Liquidity and interest rates. *Journal of Economic Theory* 50: 237–264.

Manuelli, R., and T.J. Sargent. 1988. Models of business cycles: A review essay. *Journal of Monetary Economics* 22: 523–542.

Rotemberg, J.J. 1984. A monetary equilibrium model with transactions costs. *Journal of Political Economy* 92: 40–58.

Tobin, J. 1956. The interest-elasticity of the transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Vissing-Jorgensen, A. 2002. Towards an explanation of household portfolio choice heterogeneity: Nonfinancial income and participation cost structures. Working Paper No. 8884. Cambridge, MA: NBER.

Liquidity Preference

Carlo Panico

Abstract

Keynes's notion of liquidity preference stems from the fact that he made some specific sources of demand for monetary instruments depend upon the expected variations of the interest rate, and consequently on the expected variations in the capital value of financial assets. This source of demand was considered to be *the* cause of variations in the rate of interest. Economists close to Keynes realized that in the *General Theory* he had turned the analysis of liquidity preference into a new theory of the interest rate. Robertson defended the marginalist theory, while Hicks paved the way for the 'neoclassical synthesis'.

Keywords

Cambridge equation; Central bank money; Finance motive; Hicks, J. R.; IS–LM models; Keynes, J. M.; Liquidity preference; Loanable funds; Marginalist theory of the rate of interest; Monetary instruments; Natural rate and average rate of interest; Neoclassical synthesis; Precautionary motive; Reserves–loans ratio; Robertson, D.; Saving and investment; Simultaneous equilibrium; Speculation; Speculative motive; Subjective probability; Temporary equilibrium; Tobin, J.; Transaction motive; Uncertainty; Walras's Law

JEL Classifications
E4

The notion of 'liquidity preference' has become generally used in the literature on monetary issues (particularly that concerned with the interest rate) following Keynes's contributions in the 1930s. It concerns the motives for demanding monetary instruments or other close substitutes. Earlier, the analysis of the demand for monetary instruments was based on other motives and concepts and led to different conclusions.

The analysis of the motives for demanding monetary instruments plays a specific role within monetary theory. The literature dealing with the interest rate, for instance, has always distinguished two different analytical steps. The first deals with the *variations* in the 'market interest rate', that is, that actually observed everyday: it *describes* how a change in this rate (or in the structure of the interest rates) comes about. To do that, it provides an analytical scheme which describes the behaviour of the money markets, by considering one after the other all different sources of demand for and supply of monetary instruments, pointing out the main causes of their variations. The second step deals with the *level* of the interest rate. It *explains* why this rate tends to remain, over a specific period of time, at a certain level, pointing out the factors determining it. The way in which these factors operate is then described by using the scheme provided in the first step of analysis. This clarifies the market mechanisms (that is, changes in the different components of demand and supply in the money markets) through which the prevailing level of the interest rate asserts itself.

The analysis of the motives for demanding monetary instruments thus properly belongs to the first step: it cooperates to describe the working of the money markets and the way in which variations in the interest rate (or in the structure of interest rates) occur.

This approach was followed by Smith, Ricardo, Tooke, J.S. Mill, Marx, Marshall, Wicksell, J.M. Keynes, Robertson, and so on, independently of the particular theory they proposed, that is whether the level of the 'average interest rate' (that prevailing over a specific period of time) was determined by the 'forces of productivity and thrift' or by other factors.

Prior to Keynes's contributions in the 1930s, it was assumed that monetary instruments (in most cases, central bank money) are demanded for two reasons. First, they are demanded by the household sector for the 'circulation of income'.

Households, that is, hold in the form of currency a certain fraction of their income to carry out their daily consumption expenditure.

The second source of demand for central bank money, it was assumed, comes from the banking sector which requires liquid reserves to make payments to depositors and to meet the demand for bank loans of different maturity. Banks' decisions, it was argued, are concerned with protecting themselves against the risk of running out of liquid means while minimizing cost. In such analyses, which did not use modern portfolio choice tools, the amount of reserves banks demand depends upon the composition of their portfolio (particularly the maturity of their loans) and upon the degree of uncertainty they feel as to the smooth operation of the credit payment system. On the basis of these two elements, banks fix the desired ratio between their reserves in central bank money and the amount of loans they can supply.

As some authors noticed, the presence of uncertainty among the elements affecting the decisions of financial operators makes the credit payment system unstable. The desired ratio of reserves to loans changes continuously and sometimes sharply. Financial markets become tighter precisely when more liquid means are required. A higher degree of uncertainty as to the smooth operation of the system, for instance, leads the business and the banking sectors to desire to 'become more liquid'. The former tend to discount a larger amount of bills of exchange (that is, demand more short-term bank loans), while the latter set at a higher level the desired reserves–loans ratio, so supplying a smaller amount of bank loans.

The instability of the system and the variability of the interest rates were therefore recognized by

some economists (a minority) and ascribed to the uncertainty felt by banks and business as to their ability to solve cash-flow problems.

This analysis of the demand for monetary instruments was dominant from Adam Smith onwards. Its basic points were still reflected in the famous 'Cambridge equation' presented by Pigou in his article 'The value of money' (1917) and in Keynes's *A Tract on Monetary Reform* (1923).

Keynes's Analysis of Liquidity Preference

The analysis of the motives for demanding monetary instruments was considerably refined by J.M. Keynes in the 1930s. Developing the analysis inherited from Marshall and Pigou, Keynes distinguished three motives for demanding monetary instruments (by which was now meant member banks' money, that is, bank deposits).

First, monetary instruments are demanded for transaction purposes. The amount demanded due to this motive is a stable function of the level of income.

The second source of demand for monetary instruments is for precautionary purposes, defined as the demand coming from different sectors as a protection against the possibility that some unexpected payment has to be made, or that some expected receipts cannot be realized. This definition has been differently interpreted. Some authors (and the majority of textbooks) have interpreted it in a restrictive way, by identifying it with the households' holding of bank deposits as a precaution against extraordinary events (for example, payment of hospital bills). The precautionary demand for monetary instruments was typically lumped together with the transaction demand, both being an increasing function of the level of income.

Other authors have instead given more extensive interpretation of this motive by including in it the demand coming from all financial operators feeling highly uncertain as to the future level of the interest rate. R. Kahn (1954) explained that people prefer holding part of their wealth in liquid means when their knowledge as to how the rate of interest is going to behave in the near future is so limited as to make it impossible to consider some future levels of this rate more probable than others.

This way of interpreting the precautionary motive makes it close to the third motive for demanding monetary instruments identified by Keynes: the speculative motive. Speculation in financial assets occurs because some agents expect with sufficient conviction that the rate of interest will move in a certain direction. The existence of uncertainty (that is, that lack of 'complete knowledge') is not denied. Yet the 'limited knowledge' available allows some agents to consider some future levels of the rate of interest more probable than others. Monetary instruments are so demanded (to avoid a loss in the capital value of financial assets) because a rise in the rate of interest is expected, and not because of the lack of any conviction as to the future of the rate of interest (as in the case of precautionary motive).

The novelty introduced by Keynes (some authors claim that it had been anticipated by Lavington 1921) lies not in the fact that he recognized that money is also a 'store of value' (an element already present in previous literature), but in the fact that he made some specific sources of demand for monetary instruments depend upon the expected variations of the interest rate, and consequently on the expected variations in the capital value of financial assets.

On account of its magnitude, but principally on account of its high variability, which is due to the uncertain character of expectations about future events, this latter source of demand played a central role in Keynes's writing. It was considered to be *the* cause of variations in the rate of interest. Indeed, in subsequent years, some authors even identified the notion of liquidity preference with speculative motive, while many others put it at the centre of the intense debates on interest rate after the publication of the *General Theory of Employment, Interest and Money* (1936).

Keynes's innovations stimulated many controversies dealing with different aspects of the theory of interest and money. A central point in these debates was the evaluation of Keynes's own

contribution: had he really presented a new theory of the rate of interest, alternative to the dominant marginalist one?

In the preparatory works and in the *General Theory* itself, Keynes had so characterized his contribution. He had tried to show the existence of logical inconsistency in the dominant real theory and, in opposition to it, had argued in favour of a *monetary* theory of the rate of interest based on historical and conventional factors.

The essential elements of the analysis of liquidity preference had already been introduced in *A Treatise of Money*, where the marginalist theory determining the 'natural' level of the interest rate on the basis of functions of demand for investment and supply of saving was still accepted. Here liquidity preference was integrated within the marginalist theory. In the *General Theory*, instead, the notion of a 'natural' interest rate was rejected. The 'average' level of the interest rate over a specific period of time was now determined by those factors able to affect the 'common opinion' as to the prevailing value of this rate in the future, and among these factors some importance was given to the policy of the monetary authority.

Thus, while in *A Treatise on Money* the novelty of liquidity preference referred to the first step of the analysis of the interest rate (that describing how variations in this rate come about), in the *General Theory*, the novelty regarded the second step of analysis, that is the theory determining the level of this rate.

Robertson's Critique After the General Theory

The group of economists close to Keynes in those years, with whom he discussed the proofs of the *General Theory*, fully realized that only in this book had he turned the analysis of liquidity preference, already present in the *Treatise*, into a new theory of the interest rate. Not all of them, however, agreed with him. Robertson, brought up in the same Marshallian tradition as Keynes, defended the marginalist theory, claiming that Keynes was in the *General Theory* overstating the role played by monetary factors (see Keynes 1973a, pp. 499, and Robertson 1936, 1940). He invited Keynes to attribute to monetary and real forces their proper place, as he had done in *A Treatise on Money*. The abandonment of the 'forces of productivity and thrift', when dealing with the determination of the 'average' interest rate over long periods of time, left the 'expected normal value' of this rate unexplained. Robertson could not accept that 'the common opinion' as to the future value of the interest rate should be explained in terms of factors changing from one historical period to the others, rather than by referring to one specific set of factors able to affect the course of events in different historical contexts. If we ask, Robertson stated, 'what ultimately governs the judgement of wealthowners as to why the rate of interest should be different in the future from what it is today, we are surely led straight back to the fundamental phenomena of productivity and thrift' (Robertson 1940, p. 25).

To clarify his view, Robertson translated Keynes's arguments into a different analytical framework based on 'flow' concepts. The determination of the 'market interest rate' (that actually observed daily) and of the 'average interest rate' (the one prevailing over long periods of time) was analysed in terms of 'loanable funds', to show that in both cases (but especially in the long-period case) the influence of the demand function for investment and the supply of saving could not be ignored.

Within this discussion, Robertson also pointed out the need for extra funds to finance new investment.

The debate with Robertson was intense. Other economists also joined in to discuss the three issues raised: whether Keynes's theory left the determination of the average interest rate 'hanging in the air' (or 'hanging by its own bootstraps'); the role of speculative motive and saving and investment within a 'loanable-funds' approach; the 'finance' motive.

Hicks and the Rise of the 'Neoclassical Synthesis'

While the debate with Robertson moved on the common ground of the Marshallian tradition,

those with other economists were characterized, from the beginning, by greater problems of understanding and communication.

A major figure in these debates was J.R. Hicks, whose reviews of the *General Theory* (Hicks 1936, 1937) were discussed with Keynes in an exchange of correspondence (see Keynes 1973b, pp. 71–83). This correspondence reveals Keynes's insistence on his inability to understand the meaning and the aim of Hicks's claim that the validity of Keynes's theory of interest did not prove other theories to be wrong.

Hicks's aim was to integrate Keynes's ideas within an approach, different from the Marshallian one, based on a new version of the neoclassical theory of value which used the notion of *temporary* general equilibria. The rate of interest was determined, with the other distributive variables, relative prices and the level of activity, within an analysis characterized by interdependence between different markets and the simultaneous attainment of equilibrium between supply and demand in all of them. Equilibrium between saving and investment decisions was reached simultaneously with equilibrium between supply of and demand for monetary instruments. The application of 'Walras's Law' then made it possible to argue that the claim that the rate of interest is determined in the money market and the claim that it is determined in the market for saving and investment are equivalent.

Hicks's writings had a great impact on the literature. They opened the way to the interpretation of Keynes's work known as the 'neoclassical synthesis' and to the wide use of the famous IS–LM apparatus. Indeed, orthodox 'Keynesian economics' was derived from this line of development, rather than from Keynes's own writings, as the debate on interest rate shows.

The distinction between the two steps of an analysis of the interest rate was now obscured. In spite of Keynes's explicit claim to the contrary (Keynes 1937, p. 215), the analysis of liquidity preference, which was intended as a means of describing the market mechanisms through which changes in the interest rate occur, became a theory determining the level of the interest rate. This theory was counter- posed to others – the 'loanable-funds theory' and the 'investment–saving theory' – in a long debate which in the end established what Hicks had hinted in his reviews of the *General Theory*, that is, that in a general equilibrium analysis to attribute the determination of a price or of a distributive variable to the attainment of equilibrium in one specific market makes no sense.

Now, none of the orthodox Keynesian literature mentioned any more what Keynes had emphasized: the instability of the speculative demand for money due to the uncertain character of the expectations about the future level of the interest rate. The integration of the market for monetary instruments within a general equilibrium analysis requires that the data determining the functions of demand for and supply of money have to be as stable as those determining the demand and supply functions in other markets.

The abandonment of Keynes's view of an unstable speculative demand for money was achieved by moving along two lines. First, the notion of an expected normal value of the interest rate was gradually abandoned. Second, the issue of stability was moved from a theoretical to an empirical level.

Already in *Value and Capital* (1939), Hicks had moved along the first line. After him, Modigliani (1944) derived a stable function of demand for money by referring to the risk of future increases in the interest rate, taking this risk as independent of people's specific expectations. The risk is thus *in general* low when the interest rate is high and high when the interest rate is low. Reference to specific expectations of the future value of interest rate could, instead, make the risk high when the rate is high and low when the rate is low. Finally, Tobin (1958) with the explicit aim of making the theoretical treatment of uncertainty more precise in Keynesian analysis, proposed deriving the demand function for money by including, among the data, subjective probability distributions of the future level of the interest rate, not considering any particular variation in this rate more probable than others. (The similarity with Kahn's precautionary motive mentioned above is clear.) In this analysis, stability of the demand function for money can be achieved by

adding one more assumption: any new piece of information acquired by agents does not change their subjective probability distribution. The meaning of this hypothesis is that agents have 'complete knowledge' of all relevant information, which amounts to assuming uncertainty away from the analysis. In his subsequent writings, Tobin did not return to this particular point, preferring to consider the issue of 'stability' an empirical, rather than a theoretical one. This line has been adopted by most followers of the orthodox Keynesian approach, thus avoiding complex theoretical problems. As a result, the possibility of reaching satisfactory conclusions on this issue appears more difficult.

Theories of the interest rate, which imply a departure from the dominant neoclassical tradition, whether Marshallian or modern general equilibrium versions, can also be found in the literature. They were held by authors close to Keynes during the preparation of the *General Theory*, like Joan Robinson and Kahn, and appear to reflect Keynes's original intentions more than other theories. Robinson and Kahn themselves, in subsequent years (see Robinson 1937, 1951; Kahn 1954) contributed to developing these analyses, which were also put forward by Kaldor (1939, 1970, 1982), and re-elaborated by a large group of economists, including Shackle (1967), Pasinetti (1974), Minsky (1975), Davidson (1978), Eatwell (1979), and Garegnani (1979).

Although there are some points of difference between these authors, they seem to agree on the instability of the speculative demand for money due to the uncertain character of the expectations about future level of the interest rate, and on the need to reject the neoclassical theory, for being either analytically inconsistent or for being based on the assumption of a simultaneous achievement of equilibrium in all markets, an assumption which neglects the different ways in which these markets are organized and operate.

The analyses of these authors have contributed to the development of a treatment of monetary issues which breaks with the traditional causal links between 'monetary' and 'real' variables, and where institutional elements, such as the way financial markets are organized over a certain period of time, play a central role.

These analyses make it possible to argue in favour of a 'monetary' determination of the interest rate, based on historical and conventional factors, thus supporting Robinson's claim that *any* opinion 'that is widely believed tends to verify itself, so that there is a large element of "thinking makes it so" in the determination of the interest rates' (Robinson 1951, p. 258).

The instability of the financial system and the variability of the interest rates are therefore recognized today, too, by some economists, who also allow for the influence of monetary factors on the level of activity and within the theory of value and distribution, in opposition to the dominant marginalist approach.

See Also

- ▶ Finance
- ▶ Keynes, John Maynard (1883–1946)

References

Davidson, P. 1978. *Money and the real world*. London: Macmillan.

Eatwell, J.L. 1979. Theories of value, output and employment. In *Keynes's economics and the theory of value and distribution*, ed. J. Eatwell and M. Milgate. London: Duckworth. 1983.

Garegnani, P. 1979. Notes on consumption, investment and effective demand: II. *Cambridge Journal of Economics* 3 (1): 63–82.

Hicks, J.R. 1936. Mr Keynes's theory of employment. *Economic Journal* 46: 238–253.

Hicks, J.R. 1937. Mr Keynes and the 'classics'; a suggested interpretation. *Econometrica* 5: 147–159.

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

Kahn, R.F. 1954. Some notes on liquidity preference. *Manchester School of Economics and Social Studies* 22: 229–257.

Kaldor, N. 1939. Speculation and economic stability. *Review of Economic Studies* 7: 1–27.

Kaldor, N. 1970. The new monetarism. *Lloyds Bank Review* No. 97, July, 1–18.

Kaldor, N. 1982. *The scourge of monetarism*. Oxford: Oxford University Press.

Keynes, J.M. 1923. A tract on monetary reform. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 4. London: Macmillan. 1971.
Keynes, J.M. 1930a. A treatise on money, vol. 1: The pure theory of money. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 5. London: Macmillan. 1971a.
Keynes, J.M. 1930b. A treatise on money, vol. 2: The applied theory of money. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 6. London: Macmillan. 1971b.
Keynes, J.M. 1936. The general theory of employment, interest and money. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 7. London: Macmillan. 1973a.
Keynes, J.M. 1937. Alternative theories of the rate of interest. *Economic Journal* 47: 241–252. In *The general theory and after; part II: Defence and development, collected writings of J.M. Keynes*, vol. 14, ed. D.E. Moggridge. London: Macmillan, 1973.
Keynes, J.M. 1973a. The general theory and after; part I: Preparation. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 13. London: Macmillan.
Keynes, J.M. 1973b. The general theory and after; part II: defence and development. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 14. London: Macmillan.
Keynes, J.M. 1979. The general theory and after: A supplement. In *Collected writings of J.M. Keynes*, ed. D.E. Moggridge, vol. 19. London: Macmillan.
Lavington, F. 1921. *The English capital market*. London: Methuen.
Minsky, H.P. 1975. *John Maynard Keynes*. New York: Columbia University Press.
Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88.
Pasinetti, L.L. 1974. *Growth and income distribution*. Cambridge: Cambridge University Press.
Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32: 38–65. Correction (February 1918), 209.
Robertson, D.H. 1936. Some notes on Mr Keynes' 'General Theory of Employment'. *Quarterly Journal of Economics* 51: 168–191.
Robertson, D.H. 1940. *Essays in monetary theory.* London: P.S. King.
Robinson, J.V. 1937. *Introduction to the theory of employment*. London: Macmillan.
Robinson, J.V. 1951. The rate of interest. *Econometrica* 19: 92–111. In *Collected economic papers of Joan Robinson*, vol. 2, Oxford: Blackwell, 1960.
Shackle, G.L.S. 1967. *The years of high theory.* Cambridge: Cambridge University Press.
Tobin, J. 1958. Liquidity preference as behavior risk. *Review of Economic Studies* 25: 65–86.

Liquidity Trap

Gauti B. Eggertsson

Abstract

A liquidity trap is defined as a situation in which the short-term nominal interest rate is zero. The old Keynesian literature emphasized that increasing money supply has no effect in a liquidity trap so that monetary policy is ineffective. The modern literature, in contrast, emphasizes that, even if increasing the current money supply has no effect, monetary policy is far from ineffective at zero interest rates. What is important, however, is not the current money supply but managing expectations about the future money supply in states of the world in which interest rates are positive.

Keywords

Arbitrage; Bank of Japan; Central banks; Commitment to optimal policy; Deflation bias; Euler equation; Expectations; Foreign exchange interventions; General equilibrium; Great Depression; Incentive compatibility; Inflation bias; Inflation targeting; Inflationary expectations; Keynesianism; Krugman, P.; Kydland, F.; Liquidity trap; Monetary policy; Money supply; New Keynesian Phillips curve; Nominal interest rate; Output gap; Prescott, E.; Public debt; Quantity theory of money; Real government spending; Rules versus discretion; Taylor rule; Zero bound (on shortterm nominal interest rate)

JEL Classifications

E4

A liquidity trap is defined as a situation in which the short-term nominal interest rate is zero. In this case, many argue, increasing money in circulation has no effect on either output or prices. The

liquidity trap is originally a Keynesian idea and was contrasted with the quantity theory of money, which maintains that prices and output are, roughly speaking, proportional to the money supply.

According to the Keynesian theory, money supply has its effects on prices and output through the nominal interest rate. Increasing money supply reduces the interest rate through a money demand equation. Lower interest rates stimulate output and spending. The short-term nominal interest rate, however, cannot be less than zero, based on a basic arbitrage argument: no one will lend 100 dollars unless she gets at least 100 dollars back. This is often referred to as the 'zero bound' on the short-term nominal interest rate. Hence, the Keynesian argument goes, once the money supply has been increased to a level where the short-term interest rate is zero, there will be no further effect on either output or prices, no matter by how much money supply is increased.

The ideas that underlie the liquidity trap were conceived during the Great Depression. In that period the short-term nominal interest rate was close to zero. At the beginning of 1933, for example, the short-term nominal interest rate in the United States – as measured by three-month Treasuries – was only 0.05 per cent. As the memory of the Great Depression faded and several authors challenged the liquidity trap, many economists begun to regard it as a theoretical curiosity.

The liquidity trap received much more attention again in the late 1990s with the arrival of new data. The short-term nominal interest rate in Japan collapsed to zero in the second half of the 1990s. Furthermore, the Bank of Japan (BoJ) more than doubled the monetary base through traditional and non-traditional measures to increase prices and stimulate demand. The BoJ policy of 'quantitative easing' from 2001 to 2006, for example, increased the monetary base by over 70 per cent in that period. By most accounts, however, the effect on prices was sluggish at best. (As long as five years after the beginning of quantitative easing, the changes in the CPI and the GDP deflator were still only starting to approach positive territory.)

The Modern View of the Liquidity Trap

The modern view of the liquidity trap is more subtle than the traditional Keynesian one. It relies on an intertemporal stochastic general equilibrium model whereby aggregate demand depends on current and expected future real interest rates rather than simply the current rate as in the old Keynesian models. In the modern framework, the liquidity trap arises when the zero bound on the short-term nominal interest rate prevents the central bank from fully accommodating sufficiently large deflationary shocks by interest rate cuts.

The aggregate demand relationship that underlies the model is usually expressed by a consumption Euler equation, derived from the maximization problem of a representative household. On the assumption that all output is consumed, that equation can be approximated as:

$$Y_t = E_t Y_{t+1} - \sigma\left(i_t - E_t \pi_{t+1} - r_t^e\right) \qquad (1)$$

where Y_t is the deviation of output from steady state, i_t is the short-term nominal interest rate, π_t is inflation, E_t is an expectation operator and r_t^e is an exogenous shock process (which can be due to host of factors). This equation says that current demand depends on expectations of future output (because spending depends on expected future income) and the real interest rate which is the difference between the nominal interest rate and expected future inflation (because lower real interest rates make spending today relatively cheaper than future spending). This equation can be forwarded to yield

$$Y_t = E_t Y_{T+1} - \sigma \sum_{s=t}^{T} E_t\left(i_s - \pi_{s+1} - r_s^e\right)$$

which illustrates that demand depends not only on the current short-term interest rate but on the entire expected path for future interest rates and expected inflation. Because long-term interest rates depend on expectations about current and future short-term rates, this equation can also be interpreted as saying that demand depends on long-term interest rates. Monetary policy works through the short-term nominal interest rate in the

model, and is constrained by the fact that it cannot be set below zero,

$$i_t \geq 0. \tag{2}$$

In contrast to the static Keynesian framework, monetary policy can still be effective in this model even when the current short-term nominal interest rate is zero. In order to be effective, however, expansionary monetary policy must change the public's expectations about future interest rates at the point in time when the zero bound will no longer be binding. For example, this may be the period in which the deflationary shocks are expected to subside. Thus, successful monetary easing in a liquidity trap involves committing to maintaining lower future nominal interest rates for any given price level in the future once deflationary pressures have subsided (see, for example, Reifschneider and Williams 2000; Jung et al. 2005; Eggertsson and Woodford 2003; Adam and Billi 2006).

This was the rationale for the BoJ's announcement in the autumn of 2003 that it promised to keep the interest rate low until deflationary pressures had subsided and CPI inflation was projected to be in positive territory. It also underlay the logic of the Federal Reserve announcement in mid-2003 that it would keep interest rates low for a 'considerable period'. At that time, there was some fear of deflation in the United States (the short-term interest rates reached one per cent in the spring of 2003, its lowest level since the Great Depression, and some analysts voiced fears of deflation).

There is a direct correspondence between the nominal interest rate and the money supply in the model reviewed above. There is an underlying demand equation for real money balances derived from a representative household maximization problem (like the consumption Euler equation 1). This demand equation can be expressed as a relationship between the nominal interest rate and money supply

$$\frac{M_t}{P_t} \geq L(Y_t, i_t) \tag{3}$$

where M_t is the nominal stock of money and P_t is a price level. On the assumption that both consumption and liquidity services are normal goods, this inequality says that the demand for money increases with lower interest rates and higher output. As the interest rate declines to zero, however, the demand for money is indeterminate because at that point households do not care whether they hold money or one-period riskless government bonds. The two are perfect substitutes: a government liability that has nominal value but pays no interest rate. Another way of stating the result discussed above is that a successful monetary easing (committing to lower *future* nominal interest rate for a given price level) involves committing to higher money supply *in the future* once interest rates have become positive again (see, for example, Eggertsson 2006a).

Irrelevance Results

According to the modern view outlined above, monetary policy will increase demand at zero interest rates only if it changes expectations about the future money supply or, equivalently, the path of future interest rates. The Keynesian liquidity trap is therefore only a true trap if the central bank cannot to stir expectations. There are several interesting conditions under which this is the case, so that monetary easing is ineffective. These 'irrelevance' results help explain why BoJ's increase in the monetary base in Japan through 'quantitative easing' in 2001–6 may have had a somewhat more limited effect on inflation and inflation expectations in that period than some proponents of the quantity theory of money expected.

Krugman (1998), for example, shows that at zero interest rates if the public expects the money supply in the future to revert to some constant value as soon as the interest rate is positive, quantitative easing will be ineffective. Any increase in the money supply in this case is expected to be reversed, and output and prices are unchanged.

Eggertsson and Woodford (2003) show that the same result applies if the public expects the central

bank to follow a 'Taylor rule', which may indeed summarize behaviour of a number of central banks in industrial countries. A central bank following a Taylor rule raises interest rates in response to above-target inflation and above-trend output. Conversely, unless the zero bound is binding, the central bank reduces the interest rate if inflation is below target or output is below trend (an output gap). If the public expects the central bank to follow the Taylor rule, it anticipates an interest rate hike as soon as there are inflationary pressures in excess of the implicit inflation target. If the target is perceived to be price stability, this implies that quantitative easing has no effect, because a commitment to the Taylor rule implies that any increase in the monetary base is reversed as soon as deflationary pressures subside.

Eggertsson (2006a) demonstrates that, if a central bank is discretionary, that is, unable to commit to future policy, and minimizes a standard loss function that depends on inflation and the output gap, it will also be unable to increase inflationary expectations at the zero bound, because it will always have an incentive to renege on an inflation promise or extended 'quantitative easing' in order to achieve low *ex post* inflation. This deflation bias has the same implication as the previous two irrelevance propositions, namely, that the public will expect any increase in the monetary base to be reversed as soon as deflationary pressures subside. The deflation bias can be illustrated by the aid of a few additional equations, as illustrated in the next section.

The Deflation Bias and the Optimal Commitment

The deflation bias can be illustrated by completing the model that gave rise to (1), (2) and (3). In the model prices are not flexible because firms reset their price at random intervals. This gives rise to an aggregate supply equation which is often referred to as the 'New Keynesian' Phillips curve. It can be derived from the Euler equation of the firm's maximization problem (see, for example, Woodford 2003)

$$\pi_t = \kappa\left(Y_t - Y_t^n\right) + \beta E_t \pi_{t+1} \tag{4}$$

where Y_t^n is the natural rate of output (in deviation from steady state), which is the 'hypothetical' output produced if prices were perfectly flexible, β is the discount factor of the household in the model and the parameter $\kappa > 0$ is a function of preferences and technology parameters. This equation implies that inflation can increase output above its natural level because not all firms reset their prices instantaneously.

If the government's objective is to maximize the utility of the representative household, it can be approximated by

$$\sum_{t=0}^{\infty} \beta^t \left\{ \pi_t^2 + \lambda_y \left(Y_t - Y_t^e\right)^2 \right\} \tag{5}$$

where the term Y_t^e is the target level of output. It is also referred to as the 'efficient level' or 'first-best level' of output. The standard 'inflation bias' first illustrated by Kydland and Prescott (1977) arises when the natural level of output is lower than the efficient level of output, that is, $Y_t^n < Y_t^e$.

Eggertsson (2006a) shows that there is also a deflation bias under certain circumstances. While the inflation bias is a steady state phenomenon, the deflation bias arises to temporary shocks. Consider the implied solution for the nominal interest rate when there is an inflation bias of $\overline{\pi}$. It is

$$i_t = \overline{\pi} + r_t^e.$$

This equation cannot be satisfied in the presence of sufficiently large deflationary shocks, that is, a negative r_t^e. In particular if $r_t^e < -\overline{\pi}$ this solution would imply a negative nominal interest rate. It can be shown (Eggertsson 2006a) that a discretionary policymaker will in this case set the nominal interest rate to zero but set inflation equal to the 'inflation bias' solution $\overline{\pi}$ as soon as the deflationary pressures have subsided (that is, when the shock is $r_t^e \geq -\overline{\pi}_t$). If the disturbance r_t^e is low enough, the zero bound frustrates the central bank's ability to achieve its 'inflation target' $\overline{\pi}$ which can in turn lead to excessive deflation. (While deflation and zero interest rates are due to real shocks in the literature discussed

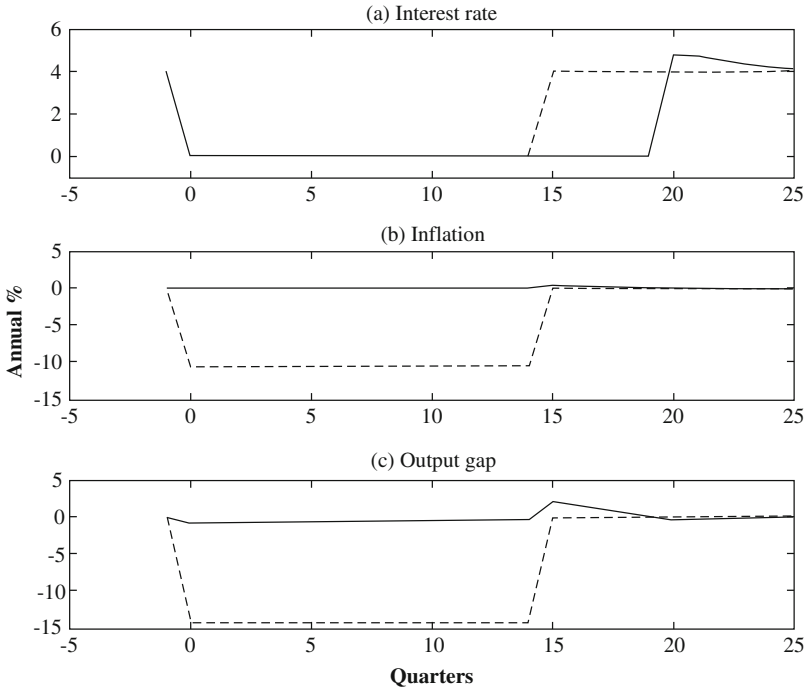
above, an alternative way of modelling the liquidity trap is that it is the result of self-fulfilling deflationary expectations; see, for example, Benhabib et al. 2001.)

To illustrate this consider the following experiment. Suppose the term r_t^e is unexpectedly negative in period 0 ($r_t^e = r_L < 0$) and then reverts back to its steady state value $\overline{r} > 0$ with a fixed probability α in every period. For simplicity assume that $\overline{\pi} = 0$. Then it is easy to verify from Eqs. (1), (4), the behaviour of the central bank described above and the assumed process for r_t^e that the solution for output and inflation is given by (see Eggertsson 2006a, for details)

$$\pi_t = \frac{1}{\alpha(1-\beta(1-\alpha)) - \sigma\kappa(1-\alpha)} \kappa\sigma r_L^e \text{ if } r_t^e = r_L^e \text{ and } \pi_t = 0 \text{ otherwise} \tag{6}$$

$$Y_t = \frac{1-\beta(1-\alpha)}{\alpha(1-\beta(1-\alpha)) - \sigma\kappa(1-\alpha)} \sigma r_L^e \text{ if } r_t^e = r_L^e \text{ and } Y_t = 0 \text{ otherwise} \tag{7}$$

Figure 1 shows the solution in a calibrated example for numerical values of the model taken from Eggertsson and Woodford (2003). (Under this calibration $\alpha = 0.1$, $\kappa = 0.02$, $\beta = 0.99$ and $r_L = -\frac{0.02}{4}$ but the model is calibrated in quarterly frequencies.) The dashed line shows the solution under the contingency that the natural rate of interest reverts to positive level in 15 periods. The inability of the central bank to set negative nominal interest rate results in a 14 per cent output collapse and 10 per cent annual deflation. The fact that in each quarter there is a 90 per cent chance of the exogenous disturbance to remaining negative for the next quarter creates the expectation of future deflation and a continued output depression, which creates even further depression and deflation. Even if the central bank lowers the short-term nominal interest rate to zero, the real rate of interest is positive, because the private sector expects deflation. The same results applies when there is an inflation bias, that is, $\overline{\pi} > 0$, but in this case the disturbance r_t^e needs to be correspondingly more negative to lead to an output collapse.



Liquidity Trap, Fig. 1 Response of the nominal interest rate, inflation and the output gap to a shocks that lasts for 15 quarters. *Note*: The *dashed line* shows the solution under policy discretion, the solid line the solution under the optimal policy commitment

The solution illustrated in Figure 1 is what Eggertsson (2006a) calls the deflation bias of monetary policy under discretion. The reason why this solution indicates a deflation bias is that the deflation and depression can largely be avoided by the correct *commitment* to optimal policy. The solid line shows the solution in the case that the central bank can commit to optimal future policy. In this case the deflation and the output contraction are largely avoided. In the optimal solution the central bank commits to keeping the nominal interest at zero for a considerable period beyond what is implied by the discretionary solution; that is, interest rates are kept at zero even if the deflationary shock r_t^e has subsided. Similarly, the central bank allows for an output boom once the deflationary shock subsides and accommodates mild inflation. Such commitment stimulates demand and reduces deflation through several channels. The expectation of future inflation lowers the real interest rate, even if the nominal interest rate cannot be reduced further, thus stimulating spending. Similarly, a commitment to lower future nominal interest rate (once the deflationary pressures have subsided) stimulates demand for the same reason. Finally, the expectation of higher future income, as manifested by the expected output boom, stimulates current spending, in accordance with the permanent income hypothesis (see Eggertsson and Woodford 2003, for the derivation underlying this figures. The optimal commitment is also derived in Jung et al. 2005, and Adam and Billi 2006, for alternative processes for the deflationary disturbance).

The discretionary solution indicates that this optimal commitment, however desirable, is not feasible if the central bank cannot commit to future policy. The discretionary policymaker is cursed by the deflation bias. To understand the logic of this curse, observe that the government's objective (5) involves minimizing deviations of inflation and output from their targets. Both these targets can be achieved at time $t = 15$ when the optimal commitment implies targeting positive inflation and generating an output boom. Hence the central bank has an incentive to renege on its previous commitment and achieve zero inflation and keep output at its optimal target. The private sector anticipates this, so that the solution under discretion is the one given in (6) and (7); this is the deflation bias of discretionary policy.

Shaping Expectations

The lesson of the irrelevance results is that monetary policy is ineffective if it cannot stir expectations. The previous section illustrated, however, that shaping expectations in the correct way can be very important for minimizing the output contraction and deflation associated with deflationary shocks. This, however, may be difficult for a government that is expected to behave in a discretionary manner. How can the correct set of expectations be generated?

Perhaps the simplest solution is for the government to make clear *announcements* about its future policy through the appropriate 'policy rule'. This was the lesson of the 'rules vs. discretion' literature started by Kydland and Prescott (1977) to solve the inflation bias, and the same logic applies here even if the nature of the 'dynamic inconsistency' that gives rise to the deflation bias is different from the standard one. To the extent that announcements about future policy are believed, they can have a very big effect. There is a large literature on the different policy rules that minimize the distortions associated with deflationary shocks. One example is found in both Eggertsson and Woodford (2003) and Wolman (2005). They show that, if the government follows a form of price level targeting, the optimal commitment solution can be closely or even completely replicated, depending on the sophistication of the targeting regime. Under the proposed policy rule the central bank commits to keep the interest rate at zero until a particular price level is hit, which happens well after the deflationary shocks have subsided.

If the central bank, and the government as a whole, has a very low level of credibility, a mere announcement of future policy intentions through a new 'policy rule' may not be sufficient. This is especially true in a deflationary environment, for at least three reasons. First, the deflation bias

implies that the government has an incentive to promise to deliver future expansion and higher inflation, and then to renege on this promise. Second, the deflationary shocks that give rise to this commitment problem are rare, and it is therefore harder for a central bank to build up a reputation for dealing with them well. Third, this problem is even further aggravated at zero interest rates because then the central bank cannot take any direct actions (that is, cutting interest rate) to show its new commitment to reflation. This has led many authors to consider other policy options for the government as a whole that make a reflation credible, that is, make the optimal commitment described in the previous section 'incentive compatible'.

Perhaps the most straightforward way to make a reflation credible is for the government to issue debt, for example by deficit spending. It is well known in the literature that government debt creates an inflationary incentive (see, for example, Calvo 1978). Suppose the government promises future inflation and in addition prints one dollar of debt. If the government later reneges on its promised inflation, the real value of this one dollar of debt will increase by the same amount. Then the government will need to raise taxes to compensate for the increase in the real debt. To the extent that taxation is costly, it will no longer be in the interest of the government to renege on its promises to inflate the price level, even after deflationary pressures have subsided in the example above. This commitment device is explored in Eggertsson (2006a), which shows that this is an effective tool to battle deflation.

Jeanne and Svensson (2007) and Eggertsson (2006a) show that foreign exchange interventions also have this effect, for very similar reasons. The reason is that foreign exchange interventions change the balance sheet of the government so that a policy of reflation is incentive compatible. The reason is that, if the government prints nominal liabilities (such as government bonds or money) and purchases foreign exchange, it will incur balance-sheet losses if it reneges on an inflation promise because this would imply an exchange rate appreciation and thus a portfolio loss.

There are many other tools in the arsenal of the government to battle deflation. Real government spending, that is, government purchases of real goods and services, can also be effective to this end (Eggertsson 2005). Perhaps the most surprising one is that policies that temporarily reduce the natural level of output, Y_t^n, can be shown to increase equilibrium output (Eggertsson 2006b). The reason is that policies that suppress the natural level of output create actual and expected reflation in the price level and this effect is strong enough to generate recovery because of the impact on real interest rates.

Conclusion: The Great Depression and the Liquidity Trap

As mentioned in the introduction, the old literature on the liquidity trap was motivated by the Great Depression. The modern literature on the liquidity trap not only sheds light on recent events in Japan and the United States (as discussed above) but also provides new insights into the US recovery from the Great Depression. This article has reviewed theoretical results that indicate that a policy of reflation can induce a substantial increase in output when there are deflationary shocks (compare the solid line and the dashed line in Fig. 1: moving from one equilibrium to the other implies a substantial increase in output). Interestingly, Franklin Delano Roosevelt (FDR) announced a policy of reflating the price level in 1933 to its pre- Depression level when he became President in 1933. To achieve reflation FDR not only announced an explicit objective of reflation but also implemented several policies which made this objective credible. These policies include all those reviewed in the previous section, such as massive deficit spending, higher real government spending, foreign exchange interventions, and even policies that reduced the natural level of output (the National Industrial Recovery Act and the Agricultural Adjustment Act: see Eggertsson 2006b, for discussion). As discussed in Eggertsson (2005, 2006b) these policies may greatly have contributed to the end of the depression. Output

increased by 39 per cent during 1933–7, with the turning point occurring immediately after FDR's inauguration, when he announced the policy objective of reflation. In 1937, however, the administration moved away from reflation and the stimulative policies that supported it – prematurely declaring victory over the depression – which helps explaining the downturn in 1937–8, when monthly industrial production fell by 30 per cent in less than a year. The recovery resumed once the administration recommitted to reflation (see Eggertsson and Puglsey 2006). The modern analysis of the liquidity trap indicates that, while zero short-term interest rates made static changes in the money supply irrelevant during this period, expectations about the future evolution of the money supply and the interest rate were key factors determining aggregate demand. Thus, recent research indicates that monetary policy was far from being ineffective during the Great Depression, but it worked mainly through expectations.

See Also

- ▶ Expectations
- ▶ Inflation Expectations
- ▶ Optimal Fiscal and Monetary Policy (with Commitment)
- ▶ Optimal Fiscal and Monetary Policy (Without Commitment)

Bibliography

Adam, K., and R. Billi. 2006. Optimal monetary policy under commitment with a zero bound on nominal interest rates. *Journal of Money, Credit and Banking* (forthcoming).

Benhabib, J., S. Schmitt-Grohe, and M. Uribe. 2001. Monetary policy and multiple equilibria. *American Economic Review* 91: 167–186.

Calvo, G. 1978. On the time consistency of optimal policy in a monetary economy. *Econometrica* 46: 1411–1428.

Eggertsson, G. 2005. Great expectations and the end of the depression. Staff report no. 234. Federal Reserve Bank of New York.

Eggertsson, G. 2006a. The deflation bias and committing to being irresponsible. *Journal of Money, Credit and Banking* 38: 283–322.

Eggertsson, G. 2006b. Was the new deal contractionary? Working paper, Federal Reserve Bank of New York.

Eggertsson, G., and M. Woodford. 2003. The zero bound on interest rates and optimal monetary policy. *Brookings Papers on Economic Activity* 2003(1): 212–219.

Eggertsson, G., and B. Pugsley. 2006. The mistake of 1937: A general equilibrium analysis. *Monetary and Economic Studies* 24(SI): 151–190.

Jeanne, O., and L. Svensson. 2007. Credible commitment to optimal escape from a liquidity trap: The role of the balance sheet of an independent central bank. *American Economic Review* 97: 474–490.

Jung, T., Y. Teranishi, and T. Watanabe. 2005. Zero bound on nominal interest rates and optimal monetary policy. *Journal of Money, Credit and Banking* 37: 813–836.

Krugman, P. 1998. It's baaack! Japan's slump and the return of the liquidity trap. *Brookings Papers on Economic Activity* 1998(2): 137–187.

Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–491.

Reifschneider, D., and J. Williams. 2000. Three lessons for monetary policy in a low inflation era. *Journal of Money, Credit and Banking* 32: 936–966.

Wolman, A. 2005. Real implications of the zero bound on nominal interest rates. *Journal of Money, Credit and Banking* 37: 273–296.

Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

List, Friedrich (1789–1846)

K. Tribe

Keywords

Economic nationalism; Infant-industry protection; International division of labour; List, F.; National economy; Productive powers; Zollverein

JEL Classifications

B31

Known chiefly as a proponent of economic nationalism and protection to 'infant industries', List's career followed a colourful, not to say disorderly course, from his engagement on behalf of

a customs union in the early 1820s to exile and residence in the United States, agitation on behalf of railway construction, energetic economic journalism, and finally to his death by suicide in November 1846, depressed by his lack of success in promoting a commercial agreement between Prussia and Britain and also by chronic financial insecurity. Born into the family of a tanner on or about 6 August 1789 in Reutlingen, Württemberg, List's early life was unremarkable. After briefly working in his father's business, he entered service in the state administration as a clerk and in 1811 secured a position in Tübingen. There he began attending the occasional law lecture, giving up his appointment in 1813 to concentrate on his legal studies. He never sat for the final lawyers' examination, instead taking and passing the actuaries' examination in September 1814.

Re-entering the administration as an accountant, he was promoted in 1816 to the position of Chief Examiner of Accounts. At the same time he became involved in the publication of a reformist journal, contributing articles on the reform of local administration. Through his connections in Stuttgart he also became involved in proposals for the creation of a new faculty for state economy at the University of Tübingen; teaching began in January 1818 and List was appointed full professor of administrative practice. List seems to have made little effort to compensate for his lack of formal academic qualification for the post, and he was dismissed in mid-1819 for absenteeism.

It is at this point that List's 'life' begins; for it transpired that his absence during April 1819 was on account of his attendance at the founding meeting of the German Association for Trade and Commerce, a body dedicated to the abolition of internal barriers to trade and which appointed List consular secretary. During the following year List travelled on behalf of the Association, and was also elected to the Württemberg representative assembly as Deputy for Reutlingen. As a result of his activities in the latter role he was tried and sentenced for sedition in 1822; appealing from the sanctuary of Baden, he failed to get the verdict altered and began a life of exile, travelling in 1823 to Paris where he made the acquaintance of Lafayette. In May 1824, believing that he had been reprieved, List returned to Stuttgart, was promptly imprisoned and, in January 1825, exiled.

Acting on a suggestion of Lafayette, List set sail for America with his family in April 1825. Taking advantage of a tour that Lafayette was undertaking at the time, List travelled and studied, making the acquaintance of several leading political figures. Settling in Pennsylvania, where he briefly tried his hand at farming, he assumed in 1826 the editorship of a German-language newspaper, the *Readinger Adler,* and became closely associated with the Pennsylvania Society for the Encouragement of Manufactures and Mechanic Arts. Through this involvement he became a supporter of the 'American system' of protective tariffs, and published in late 1827 his first serious economic work, *Outlines of American Political Economy,* which was a critique of Thomas Cooper's free-trade *Elements of Political Economy.* Such was the success of this that the Pennsylvania Society asked List to write a school textbook on political economy, but only the first chapter of this work was ever written.

As a result of an interest in coal deposits List became involved in a railway construction company which eventually opened its railroad in 1831. By this time, List had supported the presidential campaign of Jackson in 1828 and had become an American citizen; he returned to Europe, settling there permanently in late 1832 and in 1834 was appointed American Consul in Leipzig. There he became involved with the construction of the Leipzig–Dresden railway and founded the *Eisenbahnjournal* (1835), but he parted with his fellow projectors in 1837 and moved to Paris, where he spent the next three years writing a prize essay and pursuing various journalistic projects.

After his period in Paris, he moved to Augsburg and then resumed his agitation on behalf of German economic unity and south German protectionism. As before, this was largely conducted through the medium of newspapers, one of these being the *Zollvereinsblatt,* founded in 1843. These last restless years brought literary success with the publication of his *National System of Political Economy*, but little effective influence on the formation of contemporary commercial policy.

List's contribution to the formation of the Zollverein was limited to the period between 1819 and 1820, when he travelled German courts representing the cause of tariff reform. His theoretical proposals concerning protection and 'infant industries' date from his American period and are indeed a direct result of his American experience of tariff debates in the later 1820s. Much of his writing is repetitive of simple themes, as one would expect of work produced in haste for newspapers, journals and pamphlets arguing for specific reforms. However, the general logic of his position can be summarized in the following terms.

The Smithian principle of 'natural liberty' and commercial freedom was a 'cosmopolitan doctrine' which erroneously generalized the situation of Britain to the rest of the world. Commercial freedom in this sense was a freedom for Britain to dominate the world economy, thanks to the degree of development of the British economy. Free trade and economic liberty were highly desirable for a true world economy, but were only appropriate to a world of economic equals. Such a world could be created only if those countries which were in the process of development could protect their key industries against premature competition. On the international front it was necessary to create a system of treaties and agreements which would regulate trade and competition in such a way that protective tariffs and other protectionist measures would one day be redundant. On the national level, it was important to abolish internal limitations to development, such as duties between German states which hindered trade and communication. A powerful device for the creation of strong national economies was the railway, perceived not so much for its freight capacity as for its role in promoting the freedom of movement of active populations. While the abolition of internal duties opened up the fiscal geography of an economy, this space was to be given shape by a railway network which would link major centres of population – and it is this emphasis on a communications *network* that distinguishes List's work in the 1830s.

List's writing on railway development is scattered in several articles and was never presented systematically, but his conception of economic liberty and world economic development is developed in the two books he published, and the prize essay which he wrote in Paris. His *Outlines of American Political Economy* clearly contrasts a 'Smithian' economy of individuals and of mankind with 'national economy'. The error of Smith was to believe that the promotion of 'individual economy' – the satisfaction of individual wants – would lead to 'the economy of mankind' or cosmo-political economy – securing the necessities and comforts of life to the whole human race. List argued that this would not happen; the true path to the economy of mankind lay through national economy, the consideration of measures and conditions appropriate to actually existing nations. The general laws of economics outlined by Smith and his followers could manifest themselves only through these nations, which necessarily modified the operation of these laws by force of their specific 'productive powers'. The strength and independence of a national economy was secured through the control of the interior market, enabling the economy to flourish on the basis of its natural and human endowments.

The Natural System of Political Economy was written in 1837 as a response to questions concerning the ways of reconciling the interests of producers and consumers on the introduction of commercial liberty. This recapitulates the argument on individual and cosmopolitan economy already developed in the *Outlines*, but goes further in elaborating a general theory of economic development as a series of stages of agricultural, manufacturing and commercial activity. While the first stage involves a basic reliance on agriculture, by the fourth and final stage raw materials are imported for manufacture and re-export, while food is also imported.

The National System of Political Economy was published in 1841 and represents a rounding out of arguments already exposed in his earlier writing. Importantly, List now placed his arguments in a general conception of the civilizing process of international trade, underlining the fact that his opposition to free trade was by no means a narrowly nationalistic one. Also added to the original arguments is a conception of the international

division of labour elaborated on the basis of the distinction of manufacture and agriculture. List divided the world into temperate zones naturally oriented towards manufacture, and hot zones with a natural advantage in the production of agricultural goods. A balanced development of the world economy, or in other words the civilizing process, requires that the nations in the temperate zone be in equilibrium with each other and that they neither singly nor jointly exploit the lands of the hot zone, which would otherwise become dependent on manufacturing powers.

Much of the *National System* is given over to a historical account of economic development which today is very dated, while List's critique of classical economics is likewise limited by the primarily non-academic readership to which he appealed. Nonetheless, his emphasis on productive powers rather than 'value and capital', and his insistence on the specificity of national endowments and conditions in considering world economic development remain of interest. While List's primary interest lay in political and economic reform, and his audience was emergent 'informed popular opinion', he nevertheless developed conceptions of economic space and economic development that have lasting intellectual merits.

Selected Works

1827. *Outlines of American political economy.* Reprinted in *The life of Friedrich List and selection from his writings,* ed. M.E. Hirst. London: Smith, Elder & Co., 1909.

1837. *The natural system of political economy.* Trans. and ed. W.O. Henderson, London: Cass, 1983.

1841. *Das nationale System der politischen Oekonomie.* Stuttgart and Tübingen: J.G. Cotta. Trans. G.A. Matile as *National system of political economy,* Philadelphia: J.B. Lippincott & Co.

1856. Trans. S.S. Lloyd as *The national system of political economy,* London: Longmans & Co., 1885.

1927–36. *Friedrich List. Schriften, Reden, Briefe,* 10 vols. Berlin.

Litigation, Economics of

Kathryn E. Spier

Abstract

This article begins by introducing the basic economic framework for studying litigation and out-of-court settlement. One set of issues addressed is positive (or descriptive) in nature. Under what conditions will someone decide to file suit? When do cases settle out of court? Normative issues are also addressed. Are these private litigation decisions in the interest of society more broadly? Next, the article surveys some of the more active areas in the litigation literature including rules of evidence, loser-pays rules, appeals, contingent fees for attorneys, alternative dispute resolution, class actions, and plea bargaining.

Keywords

Adverse selection; Alternative dispute resolution; Asymmetric information; Bargaining; Contingent fees; Dispute resolution; English Rule; Environmental Protection Agency; Insurance contracts; Liability; Litigation; economics of; Moral hazard; Most-favourednation clauses; Negative expected value; Patents; Posner; R.; Superfund; Symmetric information; Transaction costs

JEL Classifications

K41

Litigation refers to the process of taking an argument to a court of law where a decision will be made. The discipline of economics has provided researchers – economists and legal scholars alike – with useful tools and frameworks for thinking about litigation. Is there too much litigation or too little? Why do some lawsuits go to trial while many others settle before trial? Should the losing party be required to reimburse the winning party's legal expenses? The first part of this article

presents the main frameworks for studying the economics of litigation. The second part surveys just some of the active topics in the literature.

This article is largely a condensed version of Spier (2005). Previous surveys of this topic include Cooter and Rubinfeld (1989), Hay and Spier (1998), and Daughety (2000).

Basic Framework

The Decision to Litigate

Suppose there are two *litigants*: one *plaintiff* and one *defendant*. The plaintiff is the injured party who seeks compensation; the defendant is the party who is potentially responsible for the plaintiff's injuries.

A plaintiff will rationally choose to bring suit when the expected gross return from litigation, x, exceeds the cost of pursuing the case, c_p. The gross return, x, represents the expected judgment at the end of a long and costly trial or a settlement that takes place at some time prior to the trial. It could also reflect other issues, such as the impact that a court decision will have on future cases or the plaintiff's concern for her business reputation. In general, the plaintiff's cost of pursuing the case, c_p, and the defendant's cost of fighting back, c_d, would influence the gross return, x, and could be modelled in a similar way to other economic contests (Dixit 1987). For the moment, however, we will treat them as exogenous.

The plaintiff's incentive to bring suit typically diverges from what is best for society as a whole (see Shavell 1982b, 1997). Consider a situation where accidents are totally avoided if the defendant makes a small investment in precautions. If the plaintiff were expected to sue following an accident, the defendant would rationally take the precautions. No accidents would occur and no litigation costs would be incurred. If $c_p > x$, however, then the plaintiff lacks a credible threat to sue. Knowing this, the defendant has no incentive to take the precautions (however inexpensive). In this example, the plaintiff's private incentive to sue is *socially insufficient*. This is not always the case, however. Suppose that the defendant's investment is totally ineffective: accidents occur whether or not the defendant takes precautions. Following an accident, the plaintiff will sue the defendant when $c_p < x$. The plaintiff's incentive to bring suit is *socially excessive* in this example. Litigation is a socially wasteful activity here because there is nothing the defendant could have done to avoid the accident.

Settlement

Not surprisingly, the overwhelming majority of lawsuits settle before trial. (Fewer than four per cent of civil cases that are filed in the US State Courts go to trial; see Ostrom, Kauder and La Fountain, 2001, p. 29). To use our earlier notation, the plaintiff will receive a net payoff of $x - c_p$ if the case goes to court and the defendant will receive $-x - c_d$. Although x represents a simple transfer from the defendant to the plaintiff, the litigation costs, $c_p + c_d$, represent a deadweight loss. Any out-of-court transfer $S \in (x - c_p, x + c_d)$ from the defendant to the plaintiff would be a Pareto improvement. The precise outcome of settlement negotiations will hinge on a variety of factors, including the timing of offers and counteroffers, the information and beliefs of the two litigants, and the nature of the broader legal and strategic environment.

Settlement with Symmetric Information

Suppose that the litigants are symmetrically informed and play an alternating-offer game with $T - 1$ rounds of bargaining before trial in round T. At trial, the defendant pays x to the plaintiff and the litigation costs, c_p and c_d, are incurred. The litigants share a common discount factor, δ.

This game is easily solved by backwards induction. Suppose that the plaintiff is designated to make the last settlement offer in period $T - 1$. The defendant will accept any offer that is better than going to trial, so the plaintiff will offer $S_{T-1} = \delta(x + c_d)$, minus a penny perhaps. If the case hasn't settled earlier, it will certainly settle on the courthouse steps. If we work backwards, the litigants are willing to settle for $S_{T-2} = \delta^2(x + c_d)$ in period $T - 2$, and (by an extension of this logic) are willing to settle for $S_1 = \delta_{T-1}(x + c_d)$ in period 1.

Two observations about this example are in order. First, the allocation of the bargaining surplus is sensitive to the timing of the settlement offers. If the defendant were the one to make the last offer instead, then the case would settle for $S_{T-1} = \delta(x - c_p)$ in the last round and, working backwards, we would have $S_1 = \delta^{T-1}(x - c_p)$. In other words, the party who makes the last offer succeeds in extracting all of the bargaining surplus. The bargaining surplus would, of course, be more evenly allocated in a random-offer or framework where the two litigants flip a coin to determine who makes an offer.

Second, this simple example does not predict exactly *when* settlement will take place. The litigants are, in fact, indifferent between settling for $S_1 = \delta^{T-1}(x + c_d)$ in period 1 and for $S_{T-1} = \delta(x + c_d)$ on the courthouse steps. The reason for this is straightforward: there is no inefficiency associated with delay when the litigation costs are entirely borne at trial. (Settlement models differ from the related models of bilateral trade. There, discounting causes the pie to shrink. Here, discounting by itself does not affect the size of the pie.) If the costs of litigation were incurred gradually over time instead, so the first $T - 1$ rounds of bargaining were costly as well, then there would be a unique subgame-perfect equilibrium with settlement in period *1* (Bebchuk 1996).

Settlement with Asymmetric Information

Asymmetric information is common in litigation settings. Plaintiffs often have firsthand knowledge about the damages they have suffered; defendants often have firsthand knowledge about their degree of involvement in the accident. Litigants also receive private signals concerning the credibility of their witnesses and the quality and work ethic of their lawyers. Some of this information will become commonly known over time – the parties surely learn a great deal through pretrial proceedings and discovery. Other information may never come to light at all, but can nevertheless affect trial outcomes.

Suppose that the defendant has private information about x, the expected judgment at trial. A similar analysis would follow if the plaintiff were privately informed instead. Formally, suppose x drawn from a nicely behaved probability density function $f(x)$ on $[\underline{x},\overline{x}]$ with cumulative density $F(x)$. Starting with P'ng (1983) and Bebchuk (1984), many papers assume that the uninformed player – the plaintiff in our example – makes a single take-it-or-leave-it settlement offer, S, before trial. The defendant accepts S if it is lower than what he would expect to pay at trial, $S < \delta(x + c_d)$. The offer generates a 'cut-off,' $\hat{x} = \delta^{-1}\, S - c_d$, where defendant types above the cut-off accept the offer and those below the cut-off reject the offer and go to court.

The plaintiff's optimization problem may be written as a function of the cutoff, $\hat{x}$: $Max_{\hat{x}} \int_{\underline{x}}^{\hat{x}} \delta \left(x - c_p\right) f(x)dx + [1 - F(\hat{x})]\delta(\hat{x} + c_d)$. The first term represents the plaintiff's net payoff associated with those types who reject the settlement offer, and the second term reflects the settlement payments from the defendant types above the cut-off, $\hat{x}$, who accept the offer. Any interior solution is characterized by the following first-order condition:

$$1 - F(\hat{x}) - \left(c_p + c_d\right) f(\hat{x}) = 0.$$

At least some cases will settle – the plaintiff will certainly make a settlement offer that is accepted by the most liable defendants – and an interior solution exists when $(c_p + c_d)$ is not too high.

Bebchuk's basic model has been extended in a variety of ways. Nalebuff (1987) argues that the plaintiff may no longer have a credible commitment to take the case to trial following the rejection of the settlement offer, and explicitly incorporates a credibility constraint. Spier (1992) allows the plaintiff to make a sequence of settlement offers before trial. When litigation costs are all borne at trial (so there is no efficiency loss from delay), the plaintiff waits until the very last moment to offer $S_{T-1} = \delta(\hat{x} + c_d)$, where $\hat{x}$ is defined above. (The deadline effect is less pronounced when there are pretrial costs as well.) Reinganum and Wilde (1986) let the informed litigant made a single take-it-or-leave-it offer before trial and characterize a perfect Bayesian equilibrium – unique under the D1 refinement of

Cho and Kreps (1987). The defendant's equilibrium offer $S(x) = \delta(x - c_p)$ perfectly reveals his type. Making the correct inference, the plaintiff is indifferent and accepts the settlement offer with probability

$$\pi(x) = e^{-(\bar{x}-x)/(c_p+c_d)}.$$

Note that this probability is increasing in the defendant's expected liability, x. This is implied by incentive compatibility; the defendant must be rewarded in equilibrium for making higher settlement offers with a higher rate of acceptance by the plaintiff.

Some scholars have used mechanism-design techniques to study settlement and have shown, among other things, that some cases will *necessarily* go to trial when the litigation costs are not too large (Spier 1994a). In contrast to Myerson and Satterthwaite's (1983) analysis of bilateral trade, settlement bargaining breaks down with one-sided incomplete information and despite common knowledge that gains from trade exist. (Schweizer 1989, and Daughety and Reinganum 1994, explore extensive form games with two-sided asymmetric information.) Finally, it is important to mention an older literature where litigants have different priors about the outcome at trial. Landes (1971), Posner (1973), and Gould (1973) show that settlement negotiations may fail when the two sides are sufficiently optimistic. (See Loewenstein et al. 1993, for empirical evidence on self-serving biases.)

Normative Implications

There are strong normative arguments in favour of settlement. Through a private settlement, the parties can avoid their litigation costs and (if they are risk averse) the risk premium associated with trials. *All else equal*, private settlement serves society's interest. What makes this topic more interesting – and sometimes exceptionally challenging – is that *all else is not equal*. First, settlement dilutes a defendant's incentives to avoid accidents. Following an accident, the defendant is better off if he has the option to settle his claim. Anticipating settlement on relatively advantageous terms, the defendant has less incentive to take precautions to avoid the lawsuit to begin with (Polinsky and Rubinfeld 1988). (This not necessarily a bad thing: when cases settle out of court the litigations costs are avoided so the social cost of an accident is lower. Therefore, the defendant *should* be taking less care than if all cases went to trial.) Spier (1997) shows that the defendant's incentives are diluted even further if the defendant has private information. Second, the plaintiff is made better off through settlement than she would be going to trial and is therefore more likely to bring the suit. Therefore, the anticipation of settlement raises the *overall volume of cases* that are pursued.

Topics

Accuracy

Several papers present formal analyses of the social value of accuracy in legal settings. Kaplow and Shavell (1996) argue that the *ex post* accurate verification of the victim's damages is socially valuable if the injurer knew the victim's damages at the time when he chose his precaution level. Accuracy is not valuable, however, if the victim's damages could not have been known by the injurer *ex ante*. The 'scheduling' of damages, or standardizing awards for injuries that fall into particular categories (as in workers' compensation), may be desirable in these cases. Scheduling also makes the future outcome of the case more transparent – there is less to argue about – and can help to promote settlement (Spier 1994b). Kaplow and Shavell (1992) argue that accuracy gives injurers an incentive to learn about the injuries that their activities might cause and will subsequently fine-tune their precautions. (Accurate information created by earlier trials may also help future actors fine-tune their actions; Hua and Spier 2005.)

Alternative Dispute Resolution

Alternative dispute resolution (ADR) refers to the formal and informal proceedings that help parties resolve their disputes *outside* of formal litigation. Unlike settlement, which is typically achieved by the litigants themselves (and their lawyers), ADR

proceedings often involve third parties who offer opinions and/or advice. Many of these systems are part of the court system, but many others are designed by the parties themselves (for example, ADR clauses in commercial contracts). In either case, ADR reflects the need to reduce the transaction costs of litigation and to make accurate decisions (Shavell 1994; Mnookin 1998). Farber and White's (1991) empirical study of medical malpractice claims suggests that non-binding arbitration provides an informative signal and encourages subsequent settlement. Yoon (2004) confirms this result, but finds that ADR neither reduces litigation costs nor significantly shortens the delay. The importance of this topic and the relative dearth of research – both theoretical and empirical – makes ADR a ripe topic for further investigation.

Appeals

In most legal systems, a litigant who is dissatisfied with a lower court's decision can appeal to a higher court. In Shavell (1995), appeals can be an efficient means of correcting the errors made at the lower-court level. Appeals harness the private information of the litigants themselves: an incorrectly convicted defendant is more likely to appeal an earlier ruling since the probability of reversal is higher. In this way, resources are saved relative to random auditing. (See also Spitzer and Talley 2000.) Daughety and Reinganum (2000a) consider a Bayesian model of appeals where the upper court perceives the private decision to appeal as informative and tries to rule 'correctly' given its posterior beliefs.

Bifurcation

Landes (1993) was the first to formally analyse 'bifurcated' trials where the court establishes the defendant's negligence before determining the plaintiff's damages. One benefit of bifurcation is that, once the defendant is absolved of liability, no further costs are incurred. The effect on the settlement rate is ambiguous, however. Chen et al. (1997) consider these issues in a model with asymmetric information. Daughety and Reinganum (2000b) endogenize the level of litigation spending. White (2002), in her empirical analysis of asbestos trials, shows bifurcation raises the plaintiffs' expected returns and increases the number of cases that are filed.

Case Selection

The cases that go to trial are the tip of the iceberg – the vast majority of cases are settled before trial. These tried cases are likely to differ – perhaps systematically – from the cases that never reach the courtroom. Suppose the defendant is privately informed about the expected judgment at trial. Both the screening (Bebchuk 1984) and signalling (Reinganum and Wilde 1986) approaches discussed earlier predict that defendants with weak cases are more likely to settle out of court than defendants with strong cases. Intuitively, a defendant who expects an adverse judgment is more likely to accept a settlement offer. This result would be reversed if the plaintiff has private information instead. Many authors have explored case selection using models with non-common priors instead of asymmetric information. Most notably, Priest and Klein (1984) predicted that, for tried cases, the plaintiff win rate will tend towards 50 per cent. This stark result depends on the symmetry of the litigants, among other things. (With asymmetric information, Shavell 1996, shows that any plaintiff win rate is possible.) More generally, however, the Priest–Klein framework suggests ways that trial rates may be systematically related to plaintiff win rates. Waldfogel (1995) estimates a structural model and finds results roughly consistent with the Priest–Klein theory.

Class Actions

When an injurer has harmed a group of victims, these victims may (under some circumstances) join their claims for the purpose of litigation and/or settlement. One advantage of consolidation is the scale economies associated with common proceedings and legal representation. Che (1996) assumes that plaintiffs who join a class forgo a fine-tuned award and receive instead the average damage of the group. Absent settlement, it is clear that plaintiffs with weak cases are more likely to join a class. This adverse selection problem is mitigated when plaintiffs are privately informed.

Weak plaintiffs have an incentive to remain independent, too, in an attempt to 'signal' that they have strong cases and, in equilibrium, fewer weak plaintiffs join the class. Che (2002) argues that classes may form to increase the members' bargaining power via information aggregation. The defendant is more generous when bargaining with the class as a whole than when bargaining with individuals.

Contingent Fees

In the United States, plaintiffs' attorneys are often paid on a contingent basis, receiving a third (say) of any settlement or judgment but nothing if the case is lost. The use of contingent fees is regulated in the US. In particular, lawyers are prohibited from purchasing cases from their clients (Santore and Viard 2001). Many European countries prohibit contingent fees altogether. There are many economic rationales for contingent fees. First, they give liquidity-constrained plaintiffs a way to finance their cases and shift some of the risk to the attorney. They also mitigate moral hazard (Danzon 1983) and adverse selection problems. In Rubinfeld and Scotchmer (1993), attorneys have private information about their abilities and signal high quality through a willingness to accept contingent payment. Menus of contingent fees also arise when the clients have private information. (See also the mechanism-design model of Klement and Neeman 2004.) In Dana and Spier (1993), the attorney has private information about the merits of the plaintiff's case. With contingent fees, the plaintiff can rest assured that the attorney will decline cases that are sure to lose. Finally, contingent fees can also be used strategically to make plaintiffs into 'tougher' negotiators (Hay 1997; Bebchuk and Guzman 1996). In empirical studies, Danzon and Lillard (1983) show a higher drop rate with contingent fees, and Helland and Tabarrok (2003) find that contingent fees are associated with higher-quality cases and faster case resolution.

Decoupling

It may be socially desirable to tax or subsidize the plaintiff's damage award. In Polinsky and Che (1991), a defendant chooses his level of precautions and, if injured, the plaintiff decides whether to bring suit. The optimal decoupled scheme taxes the plaintiff's award so that only a handful of cases are brought, but, at the same time, it makes the award very large so that the defendant's incentives are maintained. Since the defendant's stakes are large relative to the plaintiff's, the defendant will tend to spend more at trial (Kahan and Tuckman 1995; Choi and Sanchirico 2004). Daughety and Reinganum (2003) consider these issues in a model with asymmetric information.

Disclosure and Discovery

Litigants may voluntarily share information before trial. Indeed, the 'unravelling' logic of Grossman (1981) implies that all private information would come to light because an adverse inference would be drawn from silence. Full unravelling cannot occur, however, when hard evidence is simply unavailable. Guilty defendants have an incentive to pool with the innocent defendants who are unable to prove their innocence, for example. This suggests an important role for laws that require litigants to share information before trial. 'Discovery' can improve the accuracy of later court decisions (Hay 1994; Cooter and Rubinfeld 1994) and facilitate settlement negotiations before trial by narrowing the scope of asymmetric information (Shavell 1989). (In contrast, Schrag 1999, argues that discovery can lead to higher litigation costs and longer delays.) In Farber and White's (1991) sample of medical malpractice cases, many lawsuits are settled or dropped following discovery. Using a survey of attorneys in federal civil cases, Shepherd (1999) finds defendants increase their discovery efforts, 'tit-for-tat', in response to heightened discovery requests by the plaintiff.

The English Rule

In the United States, litigants bear their own costs of litigation – the 'American Rule'. In contrast, the 'English Rule' shifts the winner's costs to the loser. Shavell (1982a) and Katz (1990) show that the English Rule discourages the filing of lowprobability- of-prevailing cases but encourages high-probability-of-prevailing cases.

(Kaplow 1993, and Polinsky and Rubinfeld 1998, discuss the normative implications.) The English Rule also tends to raise the litigation rate when parties disagree about the probability of winning (Bebchuk 1984; Shavell 1982a). Intuitively, the scope for disagreement is even higher because the parties have different beliefs about who will bear the litigation costs. Finally, the English Rule tends to raise the level of litigation spending (Braeutigam et al. 1984; Hause 1989; Katz 1987). Intuitively, the marginal cost associated with spending is lower since the costs are partially externalized.

Inquisitorial Versus Adversarial Systems

In adversarial systems, each side gathers and processes information separately. In inquisitorial systems – such as those found in continental Europe – these activities are more centralized and often presided over by a judge (see the discussion in Parisi 2002). Adversarial systems are often criticized for giving litigants an incentive to hide relevant information from each other and from the court. They also can lead to the wasteful duplication of effort. On the other hand, adversarial systems may provide better incentives for information gathering (Dewatripont and Tirole 1999). Milgrom and Roberts (1986) present a persuasion game where the parties have equal access to all of the relevant evidence and show that accuracy is not compromised in equilibrium. This stark result may no longer hold when parties have asymmetric access to evidence or when evidence is costly to gather and disclose; see also Shin (1998), Daughety and Reinganum (2000b) and Froeb and Kobayashi (1996).

Insurance Contracts

It is common for insurance contracts to place an upper bound on the level of coverage. This creates a potential conflict between the defendant and his insurer when deciding to settle a case (Meurer 1992; Sykes 1994). The insurance company is averse to settling because the defendant will bear the downside of a very large judgment at trial. Nevertheless, the defendant may delegate settlement authority to his insurer as a strategic commitment to be 'tough' in settlement negotiations. By reducing the most that the insurer is willing to pay in settlement, the insurance contract serves to extract value from the plaintiff. These contracts may be undesirable from a social welfare perspective, however, since the toughness of the insurer can increase the litigation rate (and the associated litigation costs). Formally, these ideas are related to Aghion and Bolton's (1987) analysis of contracts as a barrier to entry. (Spier and Sykes 1998, show that corporate debt has a similar strategic value.)

Joint and Several Liability

There are many situations where a single victim is harmed by the actions of many injurers (for example, toxic-tort and price-fixing cases). Common rules for allocating responsibility include non-joint liability, where each losing defendant is responsible for his own share of damages, and joint and several liability, where a single losing defendant can be held responsible for the entirety of the plaintiff's damages. Kornhauser and Revesz (1994) analyse settlement incentives when the liability of a non-settling defendant is reduced, dollar for dollar, by the value of the previous settlements. (If the plaintiff's damages are \$80 and one defendant settles for S, the remaining defendant may be responsible for \$80 − S.) This rule encourages settlement when the cases are positively correlated but discourages settlement when the cases are independent. Some empirical support has been found in disputes between the Environmental Protection Agency (EPA) and Superfund defendants (Chang and Sigman 2000).

Most-Favoured-Nation Clauses

Settlement contracts in environments with multiple plaintiffs sometimes include 'most-favoured-nation' (MFN) clauses. They work in the following way: if an early settlement agreement includes an MFN clause and the defendant settles later with another plaintiff for more money, the early settlers receive the better terms, too. Spier (2003a) argues that MFN clauses economize on delay costs when a single defendant makes repeated offers to privately informed plaintiffs. MFNs may also be used to extract value from future plaintiffs (Spier

2003b; Daughety and Reinganum 2004). Intuitively, an MFN commits the defendant to be tough in future negotiations, allowing the defendant and the early plaintiffs to capture a greater share of the future bargaining surplus. The welfare effects of most-favoured-nation clauses are ambiguous. They can make early settlement negotiations more efficient but may lead later negotiations to fail.

Negative Expected Value Claims

Suppose that a plaintiff has a negative expected value (NEV) claim – he stands to lose money if the case proceeds all the way to trial. Could this plaintiff succeed in extracting a settlement from the defendant? Interestingly, the *divisibility* of litigation costs over time can make the plaintiff's threat to litigate the NEV claim credible (Bebchuk 1996). Here is the intuition. With divisibility, the bulk of the costs are sunk once the case reaches the courthouse steps. At that point, the plaintiff's threat to litigate is credible, so the defendant will settle. If we work backwards, the plaintiff's threat to continue may be credible at all stages of the game. Furthermore, a privately informed plaintiff with a NEV claim may mimic a plaintiff with a positive expected value claim and the defendant (not knowing for sure) may capitulate (Bebchuk 1988; Katz 1990). Finally, Rosenberg and Shavell (1985) present a model where the defendant must sink some defence costs or risk a summary judgment before trial.

Offer-of-Judgment Rules

Under Rule 68 of the United States Rules of Civil Procedure, if a plaintiff rejects a settlement offer and later receives a judgment that is less favourable, then the plaintiff is forced to bear the defendant's post-offer costs. Other rules allow for twosided cost shifting. Spier (1994a) shows that these rules raise the settlement rate when liability is acknowledged but there is private information about damages. Intuitively, the rule serves to discipline aggressive settlement tactics (but see Farmer and Pecorino 2000, and Miller 1986). Bebchuk and Chang (1999) show that offerof-judgment rules level the playing field in bargaining and lead to settlements that more accurately reflect the expected judgment at trial.

Patent Litigation

Suppose that a patentee and an imitator are trying to settle a dispute. At trial, the patent may be invalidated, in which case the imitator will compete on equal footing with the patentee. Settlement provides an opportunity for collusion. Shapiro (2003) discusses these mechanisms and proposed criteria for judicial approval of patent settlements; see also Meurer (1989). Marshall et al. (1994) argue that the mere threat of patent litigation may be enough to soften competition in a patent race; see also Choi (1998). Lanjouw and Schankerman (2001) document interesting correlations between litigation decisions and the characteristics of the patents. In particular, a patent is more likely to be litigated if it serves as the 'base of a cumulative chain' or, in other words, there are more rents to be captured from future innovators.

Plea Bargaining

In criminal cases in the United States, the prosecutor and the defendant often negotiate a guilty plea in exchange for a lighter sentence – a process known as plea bargaining. Landes (1971), in the first formal analysis of plea bargaining, assumes that the prosecutor maximizes the sum of expected sentences subject to a resource constraint. Grossman and Katz (1983) assume that the defendant privately observes his guilt and the uninformed prosecutor makes a single take-it-or-leave-it offer of a reduced sentence in exchange for a guilty plea. In the screening equilibrium, the guilty defendants accept the offer and the innocent defendants reject the offer and go to trial. This is, of course, similar to Bebchuk's (1984) analysis of civil settlement. In Reinganum (1988), the prosecutor's offer signals the prosecutor's private information and, as in Reingaum and Wilde's (1986) analysis of civil settlement, the offers with high sentences are rejected more. In contrast to Grossman and Katz (1983), trials are more likely when the defendant is guilty. (In Reinganum 2000, an informed defendant makes an offer to an uninformed prosecutor.)

Precedent

In Anglo-American legal systems, laws can be created and changed by judges over time. Cooter et al. (1979) present an early formal model where the courts learn about – and subsequently adjust – standards of care for injurers and victims. Landes and Posner (1976) consider the possibility of judicial bias, but argue that the threat of being overruled mitigates a judge's incentive to pursue his own agenda. Gennaioli and Shleifer (2005) present a formal model with a different conclusion. Rasmusen (1994) formalizes strategic interactions among a sequence of judges in a dynamic framework and shows that judges may cooperate in equilibrium and follow past precedents because violations would lead to future breakdowns where their own precedents would be violated by others; see also Schwartz (1992), Daughety and Reinganum (1999b) and Kornhauser (1992). Levy (2005) presents a model where judges have career concerns and go against precedent to signal their abilities. (A set of related rules and doctrines, 'collateral estoppel', applies when at least one litigant is involved in multiple suits; see Spurr 1991, and Che and Yi 1993.)

Secret Settlement

It is not uncommon lawsuits to settle secretly, where neither the existence of the suit nor the terms of the settlement are observed by the public. Secrecy may be facilitated through 'gag orders' or through private contracts. In Daughety and Reinganum (1999a, 2002), open settlements publicize the defendant's involvement in a case and increase the likelihood that other plaintiffs will file suit in the future. They also provide future plaintiffs with information about the expected value of their claims. Daughety and Reinganum (1999a) show that, because of the publicity effect, early plaintiffs can extract 'hush money' from defendants, enriching themselves at the expense of later plaintiffs. Importantly, secrecy can compromise firms' behaviour and product safety choices in a market setting (Daughety and Reinganum 2005).

Standards of Proof

How confident should a judge or jury be before convicting a defendant or finding in favour of a plaintiff? Rubinfeld and Sappington (1987) present a framework where the defendant can manipulate the signal received by the court, and shows how the optimal standard of proof balances litigation costs and *ex ante* deterrence concerns. Sanchirico (1997) presents a model where plaintiffs, as well as defendants, make investments in their cases. Demougin and Fluet (2006) explores the trade-offs when the defendant's wealth is limited. See Bernardo et al. (2000) and Hay and Spier (1997) for discussions of the burden of proof.

See Also

- ▶ Dispute Resolution
- ▶ Law, Economic Analysis of

Acknowledgment The author thanks the Searle Fund for financial support.

Bibliography

Aghion, P., and P. Bolton. 1987. Contracts as a barrier to entry. *American Economic Review* 77: 388–401.

Bebchuk, L. 1984. Litigation and settlement under imperfect information. *RAND Journal of Economics* 15: 404–415.

Bebchuk, L. 1988. Suing solely to extract a settlement offer. *The Journal of Legal Studies* 17: 437–450.

Bebchuk, L. 1996. A new theory concerning the credibility and success of threats to sue. *The Journal of Legal Studies* 25: 1–25.

Bebchuk, L., and H. Chang. 1999. The effect of offer-of-settlement rules on the terms of settlement. *The Journal of Legal Studies* 28: 489–513.

Bebchuk, L., and A. Guzman. 1996. How would you like to pay for that? The strategic effects of fee arrangements on settlement terms. *Harvard Negotiation Law Review* 1: 3–63.

Bernardo, A., E. Talley, and I. Welch. 2000. A theory of legal presumptions. *Journal of Law, Economics, and Organization* 16: 1–49.

Braeutigam, R., B. Owen, and J. Panzar. 1984. An economic analysis of alternative fee shifting systems. *Law and Contemporary Problems* 47: 173–204.

Chang, H.F., and H. Sigman. 2000. Incentives to settle under joint and several liability: An empirical analysis of superfund litigation. *The Journal of Legal Studies* 24: 205–236.

Che, Y.-K. 1996. Equilibrium formation of class action suits. *Journal of Public Economics* 62: 339–361.

Che, Y.-K. 2002. The economics of collective negotiations in pretrial bargaining. *International Economic Review* 43: 549–576.

Che, Y.-K., and J.G. Yi. 1993. The role of precedents in repeated litigation. *Journal of Law, Economics, and Organization* 9: 399–424.

Chen, K.-P., H.-K. Chien, and C.Y.C. Chu. 1997. Sequential versus unitary trials with asymmetric information. *The Journal of Legal Studies* 26: 239–258.

Cho, I.-K., and D. Kreps. 1987. Signalling games and stable equilibria. *Quarterly Journal of Economics* 102: 179–221.

Choi, J.P. 1998. Patent litigation as an information-transmission mechanism. *American Economic Review* 88: 1249–1263.

Choi, A.H., and C.W. Sanchirico. 2004. Should plaintiffs win what defendants lose? Litigation stakes, litigation effort, and the benefits of decoupling. *The Journal of Legal Studies* 33: 323–354.

Cooter, R., and D. Rubinfeld. 1989. Economic analysis of legal disputes and their resolution. *Journal of Economic Literature* 27: 1067–1097.

Cooter, R., and D. Rubinfeld. 1994. An economic model of legal discovery. *The Journal of Legal Studies* 23: 435–464.

Cooter, R., L. Kornhauser, and D. Lane. 1979. Liability rules, limited information, and the role of precedent. *Bell Journal of Economics* 10: 366–373.

Dana, J., and K. Spier. 1993. Expertise and contingent fees: The role of asymmetric information in attorney compensation. *Journal of Law, Economics, and Organization* 9: 349–367.

Danzon, P. 1983. Contingent fees for personal injury litigation. *Bell Journal of Economics* 14: 213–223.

Danzon, P., and L. Lillard. 1983. Settlement out of court: The disposition of medical malpractice claims. *The Journal of Legal Studies* 12: 345–378.

Daughety, A. 2000. Settlement. In *Encyclopedia of law and economics*, ed. B. Bouckaert and G. De Geest, Vol. 5. Cheltenham: Edward Elgar.

Daughety, A., and J. Reinganum. 1994. Settlement negotiations with two-sided asymmetric information: Model duality, information distribution, and efficiency. *International Review of Law and Economics* 14: 283–298.

Daughety, A., and J. Reinganum. 1999a. Hush money. *RAND Journal of Economics* 30: 661–678.

Daughety, A., and J. Reinganum. 1999b. Stampede to judgment: Persuasive influence and herding behavior by courts. *American Law and Economics Review* 1: 158–189.

Daughety, A., and J. Reinganum. 2000a. Appealing judgments. *RAND Journal of Economics* 31: 502–525.

Daughety, A., and J. Reinganum. 2000b. On the economics of trials: Adversarial process, evidence, and equilibrium bias. *Journal of Law, Economics, and Organization* 16: 365–394.

Daughety, A., and J. Reinganum. 2002. Information externalities in settlement bargaining: Confidentiality and correlated culpability. *RAND Journal of Economics* 334: 587–604.

Daughety, A., and J. Reinganum. 2003. Found money? Split-award statutes and settlement of punitive damages cases. *American Law and Economics Review* 5: 134–164.

Daughety, A., and J. Reinganum. 2004. Exploiting future settlements: A signaling model of most-favored-nation clauses in settlement bargaining. *RAND Journal of Economics* 35: 467–485.

Daughety, A., and J. Reinganum. 2005. Secrecy and safety. *American Economic Review* 95: 1074–1091.

Demougin, D., and C. Fluet. 2006. Preponderance of evidence. *European Economic Review* 50: 963–976.

Dewatripont, M., and J. Tirole. 1999. Advocates. *Journal of Political Economy* 107: 1–39.

Dixit, A. 1987. Strategic behavior in contests. *American Economic Review* 77: 891–898.

Farber, H., and M. White. 1991. Medical malpractice: An empirical examination of the litigation process. *RAND Journal of Economics* 22: 199–217.

Farmer, A., and P. Pecorino. 2000. Conditional cost shifting and the incidence of trial: Pretrial bargaining in the face of a Rule 68 offer. *American Law and Economics Review* 2: 318–340.

Froeb, L., and B. Kobayashi. 1996. Naive, biased, yet Bayesian: Can juries interpret selectively produced evidence? *Journal of Law, Economics, and Organization* 12: 257–276.

Gennaioli, N., and A. Shleifer. 2005. The evolution of precedent. Working paper no. 11265. Cambridge: NBER.

Gould, J. 1973. The economics of legal conflicts. *The Journal of Legal Studies* 2: 279–300.

Grossman, S. 1981. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics* 24: 461–483.

Grossman, G., and M. Katz. 1983. Plea bargaining and social welfare. *American Economic Review* 73: 749–757.

Hause, J. 1989. Indemnity, settlement, and litigation, or 'I'll be suing you'. *The Journal of Legal Studies* 18: 157–180.

Hay, B. 1994. Civil discovery: Its effects and optimal scope. *The Journal of Legal Studies* 23: 481–517.

Hay, B. 1997. Optimal contingent fees in a world of settlement. *The Journal of Legal Studies* 26: 259–278.

Hay, B., and K. Spier. 1997. Burdens of proof in civil litigation. *The Journal of Legal Studies* 26: 413–433.

Hay, B., and K. Spier. 1998. Settlement of litigation. In *The new Palgrave dictionary of economics and the law*, ed. P. Newman. London: Macmillan.

Helland, E., and A. Tabarrok. 2003. Contingency fees, settlement delay, and lowquality litigation: Empirical evidence from two datasets. *Journal of Law, Economics, and Organization* 19: 517–542.

Hua, X., and K. Spier. 2005. Information and externalities in sequential litigation. *Journal of Institutional and Theoretical Economics* 161: 215–232.

Kahan, M., and B. Tuckman. 1995. Special levies for punitive damages. *International Review of Law and Economics* 15: 175–185.

Kaplow, L. 1993. Shifting plaintiffs' fees versus increasing damage awards. *RAND Journal of Economics* 24: 625–630.

Kaplow, L., and S. Shavell. 1992. Private versus socially optimal provision of ex-ante legal advice. *Journal of Law, Economics, and Organization* 8: 306–320.

Kaplow, L., and S. Shavell. 1996. Accuracy in the assessment of damages. *Journal of Law and Economics* 39: 191–209.

Katz, A. 1987. Measuring the demand for litigation: Is the English Rule really cheaper? *Journal of Law, Economics, and Organization* 3: 143–176.

Katz, A. 1990. The effect of frivolous lawsuits on the settlement of litigation. *International Review of Law and Economics* 10: 3–27.

Klement, A., and Z. Neeman. 2004. Incentive structures for class action lawyers. *Journal of Law, Economics, and Organization* 20: 102–124.

Kornhauser, L. 1992. Modeling collegial courts, 2: Legal doctrine. *Journal of Law, Economics, and Organization* 8: 441–470.

Kornhauser, L., and R. Revesz. 1994. Multidefendant settlements: The impact of joint and several liability. *The Journal of Legal Studies* 23: 41–76.

Landes, W. 1971. An economic analysis of the courts. *Journal of Law and Economics* 14: 61–107.

Landes, W. 1993. Sequential versus unitary trials: An economic analysis. *The Journal of Legal Studies* 22: 99–134.

Landes, W., and R. Posner. 1976. Legal precedents: A theoretical and empirical analysis. *Journal of Law and Economics* 19: 249–307.

Lanjouw, J., and M. Schankerman. 2001. Characteristics of patent litigation: A window on competition. *RAND Journal of Economics* 32: 129–151.

Levy, G. 2005. Careerist judges. *RAND Journal of Economics* 36: 275–297.

Loewenstein, G., S. Issacharoff, C. Camerer, and L. Babcock. 1993. Self-serving assessments of fairness and pretrial bargaining. *The Journal of Legal Studies* 22: 135–158.

Marshall, R., M. Meurer, and J. Richard. 1994. Litigation settlement and collusion. *Quarterly Journal of Economics* 109: 211–239.

Meurer, M. 1989. The settlement of patent litigation. *RAND Journal of Economics* 20: 77–91.

Meurer, M. 1992. The gains from faith in an unfaithful agent: Settlement conflict between defendants and liability insurer. *Journal of Law, Economics, and Organization* 8: 502–522.

Milgrom, P., and J. Roberts. 1986. Relying on the information of interested parties. *RAND Journal of Economics* 17: 18–32.

Miller, G. 1986. An economic analysis of Rule 68. *The Journal of Legal Studies* 15: 93–125.

Miller, G. 1987. Some agency problems in settlement. *The Journal of Legal Studies* 161: 189–215.

Mnookin, R. 1998. Alternative dispute resolution. In *The new Palgrave dictionary of economics and the law*, ed. P. Newman. London: Macmillan.

Myerson, R., and M. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.

Nalebuff, B. 1987. Credible pretrial negotiation. *RAND Journal of Economics* 18: 198–210.

Ostrom, B., N. Kauder, and R. LaFountain. 2001. *Examining the work of the state courts, 1999–2000*. Williamsburg: National Center for State Courts.

P'ng, I.P.L. 1983. Strategic behavior in suit, settlement, and trial. *RAND Journal of Economics* 14: 539–550.

Parisi, F. 2002. Rent seeking through litigation: Adversarial and inquisitorial systems compared. *International Review of Law and Economics* 22: 193–216.

Polinsky, A., and Y.-K. Che. 1991. Decoupling liability: Optimal incentives for care and litigation. *RAND Journal of Economics* 22: 562–570.

Polinsky, A., and D. Rubinfeld. 1988. The deterrent effects of settlements and trials. *International Review of Law and Economics* 8: 109–116.

Polinsky, A., and D. Rubinfeld. 1998. Does the English Rule discourage lowprobability- of-prevailing plaintiffs? *The Journal of Legal Studies* 27: 519–535.

Posner, R. 1973. An economic approach to legal procedure and judicial administration. *The Journal of Legal Studies* 2: 399–458.

Priest, G., and B. Klein. 1984. The selection of disputes for litigation. *The Journal of Legal Studies* 13: 1–55.

Rasmusen, E. 1994. Judicial legitimacy as a repeated game. *Journal of Law, Economics, and Organization* 10: 63–83.

Reinganum, J. 1988. Plea bargaining and prosecutorial discretion. *American Economic Review* 78: 713–728.

Reinganum, J. 2000. Sentencing guidelines, judicial discretion, and plea bargaining. *RAND Journal of Economics* 31: 62–81.

Reinganum, J., and L. Wilde. 1986. Settlement, litigation, and the allocation of litigation costs. *RAND Journal of Economics* 17: 557–568.

Rosenberg, D., and S. Shavell. 1985. A model in which lawsuits are brought for their nuisance value. *International Review of Law and Economics* 5: 3–13.

Rubinfeld, D., and D. Sappington. 1987. Efficient awards and standards of proof in judicial proceedings. *RAND Journal of Economics* 18: 308–315.

Rubinfeld, D., and S. Scotchmer. 1993. Contingent fees for attorneys: An economic analysis. *RAND Journal of Economics* 24: 343–356.

Sanchirico, C. 1997. The burden of proof in civil litigation: A simple model of mechanism design. *International Review of Law and Economics* 17: 431–447.

Santore, R., and A.D. Viard. 2001. Legal fee restrictions, moral hazard, and attorney rights. *Journal of Law and Economics* 44: 549–572.

Schrag, J. 1999. Managerial judges: An economic analysis of the judicial management of legal discovery. *RAND Journal of Economics* 30: 305–323.

Schwartz, E. 1992. Policy, precedent, and power: A positive theory of supreme-court decision making. *Journal of Law, Economics, and Organization* 8: 219–252.

Schweizer, U. 1989. Litigation and settlement under two sided incomplete information. *Review of Economic Studies* 56: 163–177.

Shapiro, C. 2003. Antitrust limits to patent settlements. *RAND Journal of Economics* 34: 391–411.
Shavell, S. 1982a. Suit, settlement and trial: A theoretical analysis under alternative methods for the allocation of legal costs. *The Journal of Legal Studies* 11: 55–82.
Shavell, S. 1982b. The social versus the private incentive to bring suit in a costly legal system. *The Journal of Legal Studies* 11: 333–339.
Shavell, S. 1989. The sharing of information prior to settlement or litigation. *RAND Journal of Economics* 20: 183–195.
Shavell, S. 1993. Suit versus settlement when parties seek nonmonetary judgments. *The Journal of Legal Studies* 22: 1–14.
Shavell, S. 1994. Alternative dispute resolution: An economic analysis. *The Journal of Legal Studies* 24: 1–28.
Shavell, S. 1995. The appeals process as a means of error correction. *The Journal of Legal Studies* 24: 379–426.
Shavell, S. 1996. Any probability of plaintiff victory at trial is possible. *The Journal of Legal Studies* 25: 493–501.
Shavell, S. 1997. The fundamental divergence between the private and the social motive to use the legal system. *The Journal of Legal Studies* 26: 575–613.
Shepherd, G. 1999. An empirical study of the effects of pretrial discovery. *International Review of Law and Economics* 19: 245–263.
Shin, H. 1998. Adversarial and inquisitorial procedures in arbitration. *RAND Journal of Economics* 29: 378–405.
Spier, K. 1992. The dynamics of pretrial negotiation. *Review of Economic Studies* 59: 93–108.
Spier, K. 1994a. Pretrial bargaining and the design of fee-shifting rules. *RAND Journal of Economics* 25: 197–214.
Spier, K. 1994b. Settlement bargaining and the design of damage awards. *Journal of Law, Economics, and Organization* 10: 84–95.
Spier, K. 1997. A note on the divergence between the private and social motive to settle under a negligence rule. *The Journal of Legal Studies* 26: 613–623.
Spier, K. 2003a. The use of most-favored-nation clauses in settlement of litigation. *RAND Journal of Economics* 34: 78–95.
Spier, K. 2003b. Tied to the mast: Most-favored-nation clauses in settlement contracts. *The Journal of Legal Studies* 32: 91–120.
Spier, K. 2005. Litigation. In *The handbook of law and economics*, ed. A. Mitchell Polinsky and S. Shavell. Amsterdam: North-Holland.
Spier, K., and A. Sykes. 1998. Capital structure, priority rules, and the settlement of civil claims. *International Review of Law and Economics* 18: 187–200.
Spitzer, M., and E. Talley. 2000. Judicial auditing. *The Journal of Legal Studies* 24: 649–683.
Spurr, S. 1991. An economic analysis of collateral estoppel. *International Review of Law and Economics* 11: 47–61.
Sykes, A. 1994. 'Bad faith' refusal to settle by liability insurers: Some implications of the judgment-proof problem. *The Journal of Legal Studies* 23: 77–110.
Waldfogel, J. 1995. The selection hypothesis and the relationship between trial and plaintiff victory. *Journal of Political Economy* 103: 229–260.
White, M. 2002. Explaining the flood of asbestos litigation: Consolidation, bifurcation, and bouquet trials. Working paper no. 9362. Cambridge: NBER.
Yoon, A. 2004. Mandatory arbitration and civil litigation: An empirical study of medical malpractice litigation in the west. *American Law and Economics Review* 6: 95–134.

Liu, Ta-Chung (1914–1975)

Marc Nerlove

Born at Beijing, China, in 1914, died at Ithaca, New York, in 1975. Ta-Chung Liu studied civil engineering in the National Chiao Tung University (BS, 1936) and Cornell University (MCE, 1937) and later economics in Cornell (PhD, 1940). Liu served as Counselor of the Chinese Embassy (1941–6), Professor of Economics in the National Tsing-Hua University (1946–8), Economist in the International Monetary Fund, Lecturer in the Johns Hopkins University (1949–58), Professor of Economics in Cornell University (1958–75), and as Chairman of the Commission on Tax Reform (1968–70) of the Republic of China (Taiwan).

Lui's best known work deals with underidentification and structural estimation (*Econometrica*, 1960), in which he notes, in the Walrasian spirit, that everything depends on everything else; therefore, the *a priori* zero restrictions used to identify individual structural equations of a complete econometric model are, at best, suspect. It follows, if these dubious restrictions are rejected, that the structure of the economy is basically underidentified and the best we can hope for is to estimate reduced-form equations. This idea has been widely influential in recent criticism of macroeconometric modelling and responsible for the move to replace large-scale models by relatively simple vector autoregressive schemes involving only a few key

macroeconomic variables. Liu also estimated a series of models of the US economy and successively refined these models to apply to shorter and shorter time periods and, in this connection, prepared monthly estimates of US national product components.

His dissertation, published in 1946, represents the first attempt to construct national accounts for China. Concern with the statistical data of China and construction of national accounts for that country were themes which occupied him throughout his professional career, culminating in his massive study of Chinese national income and development with K.C. Yeh, published in 1965, and resulting in much testimony before the Joint Economic Committee of the Congress of the United States.

For a complete bibliography of the scientific works of T.C. Liu, see Klein et al. (1977).

Selected Works

1946a. The construction of national income tables and international comparisons of national incomes. *Studies in income and wealth* 8(4). New York: National Bureau of Economic Research, 73–118.

1946b. *China's national income, 1931–36*. Washington, DC: Brookings Institution.

1950. (With C.G. Chang.) US consumption and investment propensities: Prewar and postwar. *American economic review* 40(4): 565–582.

1954a. (With J.J. Polak.) The stability of the exchange rate mechanism in a multi-country system. *Econometrica* 22(3): 360–389.

1954b. The elasticity of US import demand: A theoretical and empirical reappraisal. *International monetary fund staff papers* 3: 416–441.

1955. A simple forecasting model of the US economy. *International monetary fund staff papers* 4: 434–466.

1960. Underidentification, structural estimation and forecasting. *Econometrica* 28(4): 855–865. Reprinted in *Selected Readings in Econometrics from Econometrica*, ed. J.W. Hooper and M. Nerlove, Cambridge, MA: MIT Press, 1970.

1963. An exploratory quarterly economic model of effective demand in the postwar US Economy. *Econometrica* 31(3): 301–348.

1965a. (With G.H. Hildebrand.) *Manufacturing production functions in the United States, 1957: An inter-industry and interstate comparison of productivity*. Ithaca: School of Industrial and Labor Relations, Cornell University.

1965b. (With K.C. Yeh.) *The Economy of the Chinese Mainland: National Income and Economic Development, 1933–59*. Princeton: Princeton University Press.

1969. A monthly recursive econometric model of the United States: A test of feasibility. *Review of economics and statistics* 51(1): 1–13.

1972. (With R.F. Engle.) Effects of aggregation over time on dynamic characteristics of an econometric model. *Economic models of cyclical behavior*, ed. B.G. Hickmkan, New York: National Bureau of Economic Research.

1974. (With E.C. Hwa.) Structure and applications of a monthly econometric model of the US. *International economic review* 15(2): 328–365.

Bibliography

Klein, L.R., M. Nerlove, and S.C. Tsiang. 1977. Ta-Chung Liu, 1914–75. *Econometrica* 45(2): 527–530.

Lloyd, William Forster (1794–1852)

Barry Gordon

Keywords

Concentration of ownership; Justice; Labour power; Lloyd, W. F.; Marginal utility theory of value; Class; Poor Law; Poverty; Subsistence wage

JEL Classifications
B31

Lloyd was Drummond Professor of Political Economy at the University of Oxford from 1832 to 1837. During those years he delivered a series of lectures which display marked originality and willingness to differ from the current canons of received wisdom among political economists. Twelve of the lectures were published. The manuscripts of the remaining lectures, approximately 24 in number, have not been found (Romano 1977).

Among the published lectures, that of 1833 and the second set of 1836 are quite outstanding. His lecture on Value (1833) has moved some leading historians of economic thought to hail Lloyd as one of the first writers to articulate the marginal utility theory of value. Less celebrated, but equally notable, is his analysis of the manner in which the operations of the contemporary British economy condemn unskilled labourers to poverty. Against the popular Malthusianism of his day, he argues in favour of the principle of poor laws and of the proposition that relief of the poor is a matter of social justice (rather than individual charity).

In the course of his 1836 lectures Lloyd constructs a model of the British economy which, he believes, demonstrates that the present situation of unskilled labourers is akin to that of slaves. Further, he observes, contemporary British society is dividing progressively into two mutually exclusive classes, and the degree of concentration of ownership and control of capital in the nation is increasing. Under existing circumstances, the unskilled worker is obliged to give ever greater quantities of his 'power of labouring' in order to obtain in return a subsistence wage.

As a person, Lloyd remains an elusive, even enigmatic, figure. He followed an older brother Charles (later, Regius Professor of Divinity and Bishop of Oxford) to Christ Church in 1812. There he studied mathematics and classics, took an MA in 1818 and was ordained in 1822. Before Lloyd succeeded Richard Whately in the Drummond Chair, he was Reader in Greek (1823) and lecturer in mathematics (1824). In 1834 he was elected a Fellow of the Royal Society. At the end of his period as Professor of Political Economy, Lloyd left Oxford to live at Prestwood, Great Missenden, Buckinghamshire, where he died in 1852. During his last 15 years Lloyd appears to have lived very quietly and published nothing. There is as yet no satisfactory explanation as to why this able and well-connected scholar chose to remain silent.

Selected Works

Twelve of Lloyd's lectures, 1834–1836, were published collectively as *Lectures on Population, Value, Poor Laws and Rent,* London, 1837; reprinted, New York: A.M. Kelley, 1968. The collection includes: Two Lectures on the Checks to Population, delivered in 1834; A Lecture on the Notion of Value as Distinguishable not only from Utility, but also from Value in Exchange, delivered in 1833; Four Lectures on Poor Laws, delivered in Hilary term, 1836; Two Lectures on the Justice of Poor-Laws, and One Lecture on Rent, delivered in Michaelmas term, 1836. Earlier, Lloyd had published *Prices of Corn in Oxford in the Beginning of the Fourteenth Century: Also from the Year 1583 to the Present Time,* Oxford, 1830.

Bibliography

Bowley, M. 1972. The predecessors of Jevons – The revolution that wasn't. *The Manchester School of Economic and Social Studies* 40(1): 9–29.

Gordon, B.J. 1966. W.F. Lloyd: A neglected contribution. *Oxford Economic Papers*, NS, 18(1): 64–70.

Harrod, R.F. 1927. An early exposition of final utility: W.F. Lloyd's lecture on the notion of value (1833) reprinted. *Economic History* (supplement to the *Economic Journal*), May, 168–183.

Romano, R.M. 1971. W.F. Lloyd – A comment. *Oxford Economic Papers* 23: 285–290.

Romano, R.M. 1977. William Forster Lloyd – A non-Ricardian? *History of Political Economy* 9: 412–441.

Loanable Funds

S. C. Tsiang

JEL Classifications
E4

The term 'loanable funds' was used by the late D.H. Robertson, the chief advocate of the loanable funds theory of the interest rate, in the sense of what Marshall used to call 'capital disposal' or 'command over capital', (Robertson 1940, p. 2). In a money-using economy where money is the only accepted means of payment, however, loanable funds are simply sums of money offered and demanded during a given period of time for immediate use at a certain price.

The loanable funds theory of interest is the theory which maintains that the interest rate, i.e. the price for the use of such funds per unit of time, must be determined by the supply and demand for such funds.

The insistence on the *flow* nature of loanable funds is based upon the crucial conception that in a money-using world the major bulk of money normally exists in a continuous circular flow. It is constantly passing out of the hands of one person as the means of payment for his expenditures into the hands of others as the embodiment of their incomes and sales proceeds, which will in turn be expended, and so on *ad infinitum*. A part of the money in this endless circular flow, however, is observed to be constantly being diverted into a side stream leading to the money market, where it constitutes the supply of loanable funds. From there borrowers of loanable funds would then take them off and in general would put them back into the main circular flow of expenditures and incomes (receipts).

This emphasis on the flow nature of loanable funds does not imply that the loanable funds theory would be unaware that there are sometimes money balances held inactive, like stagnant puddles lying off the main stream of the money flow. The loanable funds theory, however, would maintain that the stocks of money off the circular flow, as well as the stock of money inside the circular flow, have no direct influence on the money market. It is only when people attempt to divert money from the circular flow into the money market (saving), or into the stagnant puddles (hoarding), or conversely try to withdraw the inactive money from the stagnant puddles for re-injection into the circular flow or into the money market (dishoarding), that the interest rate will be directly affected. In other words, only *adjustments* in the idle balances (hoarding or dishoarding) together with the flows of savings and investment exert direct influences on the interest rate.

Since flows must be measured over time, we must choose a convenient unit to measure time. To take account of the fact that money does not circulate with infinite velocity, Robertson defined the unit period as one 'during which, at the outset of our inquiry, the stock of money changes hands once in final exchange for the constituents of the community's real income or output' (Robertson 1940, p. 65). In my opinion, however, it would be more consistent and convenient to define the unit period as one during which, at the outset, the stock of money changes hands once in exchange for all commodities and services instead of restricting the objects of exchange to final products only (Tsiang 1956, esp. pp. 545–7). The reason for this will be clear later. Based on our new definition of the unit period, all gross incomes and sales proceeds from goods and services received during the current period cannot be spent on anything until the next period when they are then said to be 'disposable'.

The definition of the unit period, however, does not preclude the funds borrowed or realized from sales of financial assets from being expendable during the same period. This differential treatment of the proceeds of sales of financial assets as distinguished from the proceeds of sales from goods and services is also an attempt to simulate the real situation in our present world; for the velocity of circulation of money against financial assets is in fact observed to be many

times faster than that against goods and services. Assuming that there is a fixed unit period in our short period analysis does not necessarily imply that we are *ipso facto* assuming the invariability of the velocity of circulation of money; for short period variations in the velocity of money can be taken care of in terms of increases or decreases in the idle balances held.

Under this definition of the unit period and the implicit assumptions behind it, each individual, therefore, faces a financial constraint in that during a given unit period he can spend only his disposable income and his idle balances (the sum of the two constitutes the entire stock of money he possesses at the beginning of the period) plus the money he can currently borrow on the money market. Buying on credit is to be treated as first borrowing the money and then spending it. Thus when he plans to spend more than his disposable income and the amount he is willing to dishoard from his idle balances, he must borrow the excess from the money market to satisfy his total demand for finance. Since additions to the demand side are equivalent to deductions from the supply side, and vice versa, we need not dispute with Robertson when he classifies the demand for, and the supply of, loanable funds on the money market as follows (Robertson 1940, p. 3).

On the demand side, he lists, with terminology slightly changed:

D1 funds required to finance current expenditures on investment of fixed or working capital;

D2 funds required to finance current expenditures on maintenance or replacement of existing fixed or working capital (note here that if our unit period were defined in the way Robertson defined it, i.e., as the period during which the total stock of money changes hands only once in the final exchange for the constituents of the community's real income, then the current expenditure on maintenance and replacement, i.e., on intermediate products, cannot be said to require a dollar for dollar provision of finance as would expenditures on final products);

D3 funds to be added to inactive balances held as liquid reserves;

D4 funds required to finance current expenditures on consumption in excess of disposable income.

Correspondingly, on the supply side, he gives:

S1 current savings defined as disposable income minus planned current consumption expenditure;

S2 current depreciation or depletion allowances for fixed and working capital taken out of the gross sales proceeds of the preceding period;

S3 dishoarding withdrawn from previously held inactive balances of money;

S4 net creation of additional money by banks.

The function of the money market is to match the flow demands for loanable funds to the flow supplies, and the instrument with which it operates to achieve equilibrium between the two sides is the vector of interest rates. It is to be noted that in the flow equilibrium condition the total stock of money does not figure at all.

Nevertheless, it must be pointed out that the flow equilibrium condition of the money market as conceived by the loanable funds theorists can imply the stock equilibrium condition as conceived by the liquidity preference theorists, provided two necessary conditions are satisfied. Of the four demands for loanable funds listed above, D1, D2 and D4 are the additional demands for transactions balances (or what Keynes in 1937 called the finance demand for liquidity) needed by some firms and consumers to finance their current planned expenditures. And of the four sources of supply of loanable funds, S1 and S2 are but the reductions in demand for finance which other consumers of firms can spare during the current period. Therefore, D1, D2 and D4 minus S1 and S2 must be equal to the net aggregate increase which the community as a whole would want to add to their transaction balances.

Similarly, D3 minus S3 is the net increase which the community would want to add to their inactive balances (including precautionary, speculative, and investment balances).

Thus the equilibrium condition of the demand for and supply of loanable funds, i.e.,

$$D1 + D2 + D3 + D4 = S1 + S2 + S3 + S4,$$

which can be rearranged as:

$$[D1 + D2 + D4 - (S1 + S2)] + (D3 - S3) = S4,$$

implies that the total increases in aggregate demand for transaction balances (finance) and for inactive balances equal the net current increases in money supply created by banks. Provided it may be presumed (*a*) that the previous stock supply of and demand for money were originally equal to each other, and (*b*) that the current increases (or decreases) in supply and demand for money (treated above as flow supply and demand for loanable funds) represent the full unlagged adjustments of the previous stock supply and demand to their new equilibrium values, the flow equilibrium of the loanable funds should necessarily imply a new stock equilibrium (Tsiang 1982).

The two necessary provisos used to be taken for granted by the liquidity preference theorists, who generally think that full stock equilibrium can be achieved instantaneously at any point in time. However, Professor James Tobin, in his Nobel lecture given in 1981 (Tobin 1982), has come to recognize that the money market cannot operate within a dimensionless point of time, but must operate in finite time periods, which he called slices of time. Furthermore, he recognized that the equilibrium which can be expected in such a short slice of time can only be that between the adjustments in the stock demanded and in the stock supplied during the period. Since adjustments in stocks per time period are flows, Tobin's new approach is thus really a sort of flow equilibrium analysis.

Moreover, Tobin, at the same time, also admitted that in such a short period as a slice of time, portfolios of individual agents cannot adjust fully to new market information. Lags in response are inevitable and rational in view of the costs of transactions and decisions. Thus neither of the two necessary conditions is satisfied in the real world. Consequently, even when the money market has brought the flow demand for and supply of loanable funds to equality, the stock demand for money and the total money stock need not have reached mutual equilibrium, which the Keynesians and the stock-approach economists used to assume as being attainable at every point of time.

Finally, it should be realized that the demand for finance for planned investment expenditure, which Keynes (1937, p. 667) admitted he should not have overlooked in his *General Theory*, is of the nature of a flow generated by a flow decision to invest. It is not just a partial adjustment of the stock demand for money towards its new equilibrium value as treated in Tobin's new theory (Tobin 1982). As Keynes put it in his reply to Ohlin (1937)>, "'Finance" is a revolving fund . . . . As soon as it is used in the sense of being expended, the lack of liquidity is automatically made good and the readiness to become temporarily unliquid is available to be used over again' (Keynes 1937, p. 666). This is essentially a reaffirmation of the traditional conception of the circular flow of money, which loanable funds theorists had emphasized from the outset, but which Keynes himself had pushed into the dark background with his emphasis that the entire stock of money is being held voluntarily in portfolio allocation.

The rediscovery of the demand for finance by Keynes and the more recent unheralded switch on the part of Tobin towards the flow approach from his usual stock approach indicate that the loanable funds theory is perhaps the more appropriate approach at least for short period dynamic analysis.

See Also

▶ Liquidity Preference

Bibliography

Keynes, J.M. 1937. The ex-ante theory of the rate of interest. *Economic Journal* 47 (December): 663–669.

Ohlin, B. 1937a. Some notes on the Stockholm theory of savings and investment, I. *Economic Journal* 47: 53–69.

Ohlin, B. 1937b. Some notes on the Stockholm theory of savings and investment, II. *Economic Journal* 47: 221–240.
Robertson, D.H. 1940. *Essays in monetary theory.* London: P.S. King.
Tobin, J. 1982. Money and finance in the macroeconomic process. *Journal of Money, Credit and Banking* 14 : 171–204.May
Tsiang, S.C. 1956. Liquidity preference and loanable funds theories, multiplier and velocity analysis: A synthesis. *American Economic Review* 46 (September): 539–564.
Tsiang, S.C. 1982. Stock or portfolio approach to monetary theory and the neo-Keynesian school of James Tobin. *IHS-Journal* 6: 149–171.

Local Public Finance

John M. Quigley

Abstract

The mobility of consumers and producers in response to fiscal incentives gives the study of local public finance its distinctive character. Households and firms are partitioned into spatial units on the basis of preferences, costs and the incentives provided by local tax and expenditure policies. These fiscal incentives are, in turn, chosen by the members of each of these jurisdictions or clubs. Externalities within and between these localities greatly affect the efficiency of taxation and the provision of public goods and services.

Keywords

Clubs; Congestion; Efficient allocation; Excise taxes; Exclusionary zoning; Intergovernmental grants; Inter-jurisdictional competition; Lindahl tax structure; Local public finance; Local public goods; Lump-sum taxes; Marginal rate of substitution; Marginal rate of transformation; Median voter theorem; Poll tax; Property taxation; Public finance; Public goods; Residential mobility; Revenue sharing; Stabilization; Tax price; Technical change; Tiebout hypothesis

JEL Classifications

H7

Economic analysis of the taxation and expenditure policies of local public authorities has become far more sophisticated as theoretical enquiry has directed attention towards the uniquely local aspects of public finance and as national policies have increased the importance of the local public sector.

Many of the issues that arise in the analysis of the local public sector are familiar reflections of the important questions in public finance that have been addressed at the national level; for example, the incidence of taxation and the welfare losses from revenue instruments; the effect of government expenditures on consumer welfare and the distribution of well-being; the effect of public sector distortions on resource allocation and relative prices.

However, the principal difference between the economic analysis of public finance at the national and at the local levels is the potential for mobility among jurisdictions by the transport of final products and inputs, and especially by residents who finance local government and consume public output. Critically, this mobility may be endogenous to the revenue or expenditure actions taken by the local public authority, and this must be considered in any economic analysis of local finance.

This insight, as it affects efficiency in the allocation of local public output and the incidence of local taxes, goes back at least to the fifth edition of Marshall's *Principles* (1907, Appendix G). Marshall presented a lucid discussion of the effect of local public expenditures on residential mobility ('A high rate spent on providing good primary and secondary schools may attract artisan residents while repelling the well-to-do' – Marshall 1920, p. 794). He also noted the effects of mobility upon the incidence of local taxes.

Given the increased complexity of decentralized taxation and expenditure patterns when compared to national government policies, one may begin by asking which economic functions of government ought to be undertaken by the

central (national) government rather than by local authorities. Consider the original Musgrave (1959) taxonomy of public sector functions: distribution, stabilization and allocation. It seems clear that a system of local taxes and expenditures is inappropriate for achieving distributional or stabilization goals. After the adoption of any system of taxation and redistribution by a locality, even one which reflects a unanimous view of the citizens, it will be in the interests of those bearing the burden of the tax to relocate in other jurisdictions and in the interests of potential beneficiaries of the redistribution to move into the jurisdiction. Similarly, locally adopted monetary and fiscal policies are unlikely to further stabilization objectives, even if such objectives are uniformly held by local citizens. Import leakages are so large that the local benefits of stabilization policies (for example, local public employment programmes) are almost certain to be less than their costs.

It is precisely the mobility of households, goods and factors across jurisdictions that defeats local stabilization and redistribution policies. Conversely, however, the same 'openness' of the local economy means that the decentralized local provision of public goods will in many cases improve the allocative efficiency of the economy. In particular, the smaller and more homogeneous a community in a system of local government, the more likely is it that the provision of public goods by any community will be consistent with the demands of its citizens. In the limit, of course, if public goods are financed by a head tax, and if there are neither economies of scale in production nor externalities in consumption, then provision by a system of small jurisdictions, each with citizens of homogeneous tastes and incomes, will result in an efficient allocation.

If, however, there are economies of scale in production, it makes sense to have larger jurisdictions. But when the public good is produced by a larger entity, 'congestion' may result; that is, the quality of the good may decline as it is shared with more people. In larger jurisdictions, moreover, citizen demands may be more heterogeneous. The problem of balancing the benefits of cost-sharing in production, on the one hand, with the sacrifice in well-being by compromising individual consumers' demands or by introducing 'congestion' in public goods consumption, on the other, has been central to the normative analysis of the local provision of public goods.

Consider, for example, a 'club' providing some collective benefit to identical individuals (Buchanan 1965). Suppose an organization supplies some public output Q subject to congestion, or equivalently, suppose it supplies a good whose standardized cost $C(N)$ increases with population N. Individuals of income Y are assessed the average cost of service provision and allocate their remaining income to some numeraire good X. A community of N identical individuals will choose public output to maximize utility, $U(Q, X)$, subject to the individual budget constraint, $Y = X + [C(N) / N]Q$. This implies the familiar Samuelson (1954) condition:

$$N[(\partial U/\partial Q)/(\partial U/\partial X)] = C(\mathrm{N}). \qquad (1)$$

The level of public good provision is chosen by the club of fixed size N so that the *sum* of the individual marginal rates of substitution (MRS) between private and public goods equals the marginal rate of transformation (MRT) in production. Given this level of public output, from the budget constraint it also follows that choice of club size to maximize utility is:

$$C'(N) = C(N)/N. \qquad (2)$$

The optimum size of the club is the membership at which the average cost of public output is equal to the marginal cost of adding another member. From equations (1) and (2) it follows that for a pure public good, that is, $C'(N)=0$, the optimal size of the club is unbounded, while for a private good, where $C(N)=PN$, the *individual* MRS is equal to the MRT and the size of the club is indeterminate.

Applied to local public finance, the model indicates that a system of communities, each with identical individuals and of that size which minimizes average cost, would be a stable and efficient mechanism for public service provision. Homogeneity of demands is necessary for efficiency even if the tax structure (or club dues) is

of the Lindahl variety. Each group in a heterogeneous community would be better off by moving to a jurisdiction with identical tax shares.

Theoretical analyses of local public economies are much more complicated when the partitioning of individuals into political jurisdictions is 'non-anonymous', that is, when the characteristics of the other members (in addition to their incomes) matter to those in the club. In many cases, an equilibrium allocation of residents to jurisdictions may not exist at all (Scotchmer 1997). As noted below, non-anonymous crowding may also affect the costs of public goods provision and the interpretation of demands for local public goods.

The 'club' model of the provision of local public goods is a special case of the so-called Tiebout (1956) model, probably the most influential idea in the modern analysis of local public finance. Tiebout's stylized and informal analysis assumes that residential mobility is costless, that local jurisdictions provide public goods at minimum average cost and that local government is financed by non-distortionary lump-sum taxes. Under these circumstances, Tiebout argues that the provision of public goods by a system of competitive local governments may be no less efficient than the allocation of private goods by the market economy. The conclusion of this argument also depends crucially upon the availability to citizens of a sufficiently large number of jurisdictions offering differing packages of local public goods and upon the absence of inter-jurisdictional externalities, as well as more conventional assumptions about full information. In reality, in most metropolitan areas, local public output is supplied by a small number of communities (small, at least, relative to the number of types of demanders); local mobility is quite costly and is motivated by many non-fiscal concerns. Individuals often live in one jurisdiction and work in another, and there are externalities among jurisdictions. Finally, revenues are raised, not by head taxes but by a variety of local levies, especially *ad valorem* taxes on real property. Each of these factors limits the economic efficiency of the local public sector in important ways.

The externalities or 'spillouts' of the benefits of public service provision mean that such goods will be underprovided without coordination by local communities – since each community will only consider the benefits accruing to its own citizens in choosing the level of service provision. For public goods and services with substantial spillouts of benefits, efficient levels of production can be stimulated by a system of open-ended matching grants to localities by the central government. As Pigou (1932) originally demonstrated, if the matching rate (the fraction of local spending reimbursed by higher government) corresponds to the fraction of local public output, which spills out to non-residents, then the externality will be internalized. It is, of course, rather difficult to implement this maxim of local public finance (Oates 1972).

The heavy reliance upon local property taxes for financing the local public sector, especially in Britain, Canada and the United States, is another source of allocative inefficiency in local finance. Clearly, a property tax alters the housing consumption decision and leads to underconsumption of housing as well as to inefficiency in public goods consumption. Until rather recently the system of local property taxes was viewed as a system of excises (Netzer 1966), regressive levies on property and housing consumption, in contrast to the original Henry George (1879) position on land taxes. Modern theoretical analyses (following Mieszkowski 1972), which assume that capital is mobile across jurisdictions and that the supply of capital is insensitive to its rate of return, have led to a reconsideration of the regressive nature of the tax. The inelastic supply of aggregate capital means that a national system of local property taxes will reduce returns to capitalists by the average level of the tax. The geographical mobility of capital implies that capital will flee from high-tax jurisdictions, raising marginal productivity and pre-tax returns, to low-tax jurisdictions, depressing pre-tax returns. Thus the incidence of the system of property taxes depends upon the magnitude of the average level of the tax, relative to the deviations from that average, as well as distribution of households among high-tax and low-tax jurisdictions. Despite the ambiguities in resolving these detailed empirical issues, this theoretical argument suggests that the burden of property

taxation is heavily skewed towards the owners of capital. Empirically, this conclusion is probably modified by regressive appraisal and administrative procedures. It should be noted, moreover, that from local governments' perspective an increase in the level of the property tax to finance service provision is an excise on property users (since a change in any one community's property tax rate can have only a negligible effect on the average level of rates for the nation).

The distortion inherent in property tax financing may lead to local policies of exclusionary zoning. If, for example, the benefits of the local public sector were roughly equal per household, then it would be in the interests of current residents to force incoming households to consume more housing than the average household. Current residents may attempt to enforce this by imposing minimum lot-size restrictions or by other exclusionary practices to increase the housing consumption of newcomers. Of course, as noted before, unless there are sufficient communities so that the households residing within a jurisdiction are literally identical, those who chose to consume less housing will typically enjoy a fiscal residual.

Despite these clear examples of allocative inefficiency in the system of local public finance and service provision, there is a substantial body of evidence that variations in property tax rates are reflected in property values and that variations in public services (for example, school quality) are capitalized into the sale prices of residential property. These findings are certainly consistent with the process of 'voting with one's feet' implied by the Tiebout model, but the capitalization of taxes and services is not necessary to efficiency in local government, nor does efficient service provision necessarily imply capitalization.

The observation that individuals register their demands for publicly financed services in their choices of community has other important implications, however. Specifically, information about the public goods provided by different jurisdictions, together with information about the characteristics of the residents of those jurisdictions, may be sufficient to identify consumer demands for public services. Extensive analyses of these issues have been undertaken, combining economic theories of the local political process with aggregate data on local public finance and choice of output. Under rather restrictive assumptions, the political process which determines the level of service provision can be modelled as the choice of the median voter of the community. Given the characteristics of that individual (or rather, estimates obtained from aggregate information), the 'tax price' that individual confronts, and the level of public output chosen, the parameters of the demand curve are estimated econometrically. The 'tax price' is the marginal cost to the individual of purchasing an additional dollar of public output. With property tax financing, this is typically approximated by the median voter's house value as a fraction of the community's taxable real property per household.

As noted above, the residents of localities may 'care' about the characteristics of other residents simply because their characteristics affect the cost of producing public services. One example may involve local schools, which absorb the largest share of local government spending on public services. To the extent that peers 'matter' in the production of educational outputs in primary school, policies of matching grants to local governments based on disadvantaged residents are called for (see Nechyba 2003). The specification of empirical models of the demand for local public services is much more problematic when the demographic characteristics reflect either tastes for public goods or the costs of supplying them, or both.

Nevertheless, the results of these empirical investigations have proven useful in the positive analysis of citizen demands for public services and in the analysis of local finance. Nevertheless, the underlying economic model of local government behaviour is open to questions, both technical (for example, the requirement that preferences exhibit single peakedness) and substantive (for example, the neglect of the role of bureaucracy in government decisions). For example, if the median voter determines the demand for local public output, then the propensity for a community to spend out of lump-sum aid from higher government ought to be no different from the

propensity to spend out of income generated by local taxation. Yet empirical evidence suggests that the propensity of communities to spend out of untied grant income greatly exceeds the propensity to spend out of ordinary income. A variety of alternative models of local finance have been espoused to help explain this 'flypaper' effect ('money sticks where it lands') in the context of bureaucratic decision-making. Chief among them are the so-called Leviathan models of a government that exploits its citizens by maximizing revenues extracted by taxation (Brennan and Buchanan 1980). Clearly, however, more theoretical work needs to be done to resolve the contradictions between mobile consumers of local public output and sluggish suppliers.

Finally, it has been suggested that the inherent nature of local output and the traditional financing mechanisms of local government combine to exacerbate the economic and administrative problems of the local public sector (Baumol 1967). Local output consists largely of labour-intensive services, where technical change is inherently slow, and is typically financed by income-inelastic tax instruments. Under reasonable demand conditions, these may produce a more or less continuous 'crisis' in local public finance, as service costs escalate more rapidly than revenue increments. Given these characteristics of the local financing mechanism, as well as the redistributive nature of many local services, there may thus be a strong case for revenue or tax-base sharing at the national level.

See Also

- ▶ Fiscal Federalism
- ▶ Public Finance
- ▶ Public Goods
- ▶ Tiebout Hypothesis
- ▶ Urban Economics

References

Baumol, W.J. 1967. Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review* 62: 415–426.

Brennan, G., and J.M. Buchanan. 1980. *The power to tax: Analytical foundations of a fiscal constitution*. Cambridge: Cambridge University Press.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* 32: 1–14.

George, H. 1879. *Progress and Poverty*. New York: Appleton.

Inman, R.P. 1979. The fiscal performance of local governments: An interpretive review. In *Current issues in urban economics*, ed. P. Mieszkowszki and M. Straszheim. Baltimore: Johns Hopkins University Press.

Marshall, A. 1907. *Principles of economics*. 5th, 8th ed. London: Macmillan, 1920.

Mieszkowski, P. 1972. The property tax: An excise tax or a property tax? *Journal of Public Economics* 1: 73–96.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

Nechyba, T. 2003. School finance, spatial income segregation, and the nature of communities. *Journal of Urban Economics* 54: 61–88.

Netzer, D. 1966. *Economics of the property tax*. Washington, DC: Brookings Institution.

Oates, W.E. 1972. *Fiscal federalism*. New York: Harcourt Brace Jovanovich.

Pigou, A.C. 1932. *The economics of welfare*. 4th ed. London: Macmillan.

Rubinfeld, D.L. 1985. The economics of the local public sector. In *Handbook of public economics*, ed. J. Auerbach and M. Feldstein, vol. 2. Amsterdam: North-Holland.

Samuelson, P.A. 1954. The pure theory of public expenditure. *Review of Economics and Statistics* 36: 387–389.

Scotchmer, S. 1997. On price-taking equilibria in club economies with nonanonymous crowding. *Journal of Public Economics* 65: 75–87.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Local Regression Models

Oliver B. Linton

Abstract

This article discusses local regression models, that is, regression models where the parameters are allowed to vary with some covariates either in a completely unrestricted fashion or in an intermediate way with some exclusion restrictions that make some parameters vary only with some covariates. Special cases are nonparametric regression and additively separable nonparametric regression.

Keywords

Additive models; Cobb–Douglas parametric model; Conditional expectation; Conditional variance; GARCH models; Generalized additive models; Identification; Linear models; Local regression models; Parametric models

JEL Classifications

C14

Local regression models are regression models where the parameters are 'localized', that is, they are allowed to vary with some or all of the covariates in a general way. Suppose that (Y, X) are random variables and let

$$E(Y|X = x) = m(x) \tag{1}$$

when it exists. The regression function $m(x)$ is of primary interest because it describes how X affects Y. One may also be interested in derivatives of m or averages thereof or in derived quantities like conditional variance $var(Y|X = x) = E(Y^2|X = x) - E^2(Y|X = x)$. In cases of heavy-tailed distributions, the conditional expectation may not exist, in which case one may instead work with other location functionals like trimmed mean or median. The conditional expectation is particularly easy to deal with but a lot of what is done for the mean can also be done for the median or other quantities.

A parametric regression model for $m(x)$ is a family of functions $M(x;\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$, where for each θ, $M(,;\theta)$ is a known function. The true parameter θ_0 for which $M(x; \theta_0) = m(x)$ for all $x \in \mathscr{X}$ is unknown and has to be estimated from data. For example, $M(x; \theta) = x^\top\theta$ would correspond to the linear regression case, which is the central model of econometrics. A key concept is that of identifiability: M is identifiable when distinct parameter values lead to different values of M for at least some x values. See Rothenberg (1971) for discussion. Parametric models arise frequently in economics and are of central importance. However, such models arise only when one has imposed specific functional forms on utility or production functions. Without these ad hoc assumptions one only gets much milder restrictions on functional form like concavity, symmetry, homogeneity and so on. The nonparametric approach is based on the belief that parametric models are usually mis-specified and may result in incorrect inferences. In this approach one treats the regression function $m(x)$ as being of unknown functional form. One usually assumes that m is a continuous function or even differentiable, although there are cases of interest where $m(x)$ is, say, continuous only from the right (left) with limits on the left (right), that is, there may be jumps at certain known or unknown locations in the support $\mathscr{X}$ of X (see Delgado and Hidalgo 2000). By not restricting the functional form one obtains valid inferences for a much larger range of circumstances. In practice, the applicability depends on the sample size and the quality of data available. The theory and methods for carrying out such estimation are well understood, and are reviewed elsewhere (Härdle and Linton 1994). Local regression models are one way of interpreting the nonparametric approach.

A local regression model is a family of functions

$$M(x; \theta(x)), \qquad \theta \in \Theta = \{\theta : \mathscr{X} \to \mathbb{R}^p\}, \tag{2}$$

where $M(x;\theta)$ is a known function of both arguments. The true (functional) parameter $\theta_0(\cdot)$ for which $M(x; \theta_0(x)) = m(x)$ for all $x \in \mathscr{X}$ is unknown. It is usually assumed to be smooth. In other words this is a standard parametric regression model except that the parameters vary with the covariate value. There are a number of special cases. At one extreme lies the parametric model in which $\theta(x) = \theta$ for all $\mathscr{X} \subset \mathbb{R}^d$, but the true θ_0 is unknown. At the other extreme lies the fully nonparametric case where $\theta(\cdot)$ is not subject to any exclusion restrictions.

Many different M functions will generally do. For example, the local constant case corresponds to $M(x; \theta) = \theta$ and the local linear case corresponds to $M(x; \theta) = \theta_0 + \theta_1 x$. These cases along with higher-order polynomials have been widely studied (see, for example, Fan and Gijbels 1996). There are also other possibilities. Consider the Cobb–Douglas parametric model

$$M(x;\theta) = \theta_0 x_1^{\theta_1} \cdots x_d^{\theta_d}, \quad (3)$$

which is widely used in studies of production. By making $\theta = (\theta_0, \theta_1, \ldots, \theta_d)$ vary freely with x one can match with any function $m(x)$ so long as the supports coincide (see, for example, Charnes et al. 1976). For binary data where it is known that $m(x) \in [0, 1]$ it is appropriate to take $M(x; \theta) = F(\theta_0 + \theta_1 x)$ for some given c.d.f. F like the normal or logit. In that case, for a given x, there exists $\theta_0(x)$, $\theta_1(x)$ such that $m(x) = F(\theta_0(x) + \theta_1(x)x)$. This example illustrates some pitfalls; for example, when $m(x) > 1$ for some x of interest. In that case, taking $M(x; \theta) = F(\theta_0 + \theta_1 x)$ will not be satisfactory.

The statistical justification for using local constant, local linear, and more generally local polynomial models is that any smooth function $m(x)$ can be approximated near the point x_0 by Taylor series expansions, so for p-times continuously differentiable scalar functions we have

$$m(x) = \sum_{j=0}^{p} \frac{1}{j!} \frac{d^j m}{dx^j}(x_0)(x - x_0)^j + R(x, x_0), \quad (4)$$

where the remainder term satisfies $R(x, x_0)/|x - x_0|^p \to 0$ as $x \to x_0$. Thus the function m is locally well approximated by a polynomial of order p, $\sum_{j=0}^{p} \alpha_j (x - x_0)^j$, where α_j can be identified with $j!^{-1} d_j m(x_0)/dx_j$. This justifies using local polynomial regression. But why should one ever work with local regression models outside the local polynomial class? First, any other local parametric model $M(x, \theta)$ that is p-times continuously differentiable in x at x_0, satisfies a similar expansion to (4), $\sum_{j=0}^{p} \beta_j (x - x_0)^j$, where β_j are functions of θ. By equating coefficients one obtains the same leading terms as long as there are 'enough' parameters in θ. Therefore, the same approximating objectives are reached by any such model. In some cases other equivalent classes may provide better approximations. Polynomials can sometimes violate some known features, like for example $m(x) \in [0, 1]$. In that case, taking $M(x;\theta)$ to be a c.d.f. of a polynomial provides the same approximation (so long as the c.d.f. chosen is also smooth enough) but imposes the boundedness restriction. Second, the local parameters may also be of interest in themselves. In the Cobb–Douglas case, the $\theta_j(x)$ can be interpreted as local elasticities. A third benefit is that the local model nests the parametric model. This leads to better statistical properties for estimators and test statistics when the model is true or approximately true, the 'home turf' case (see Hjort and Glad 1995). When the default parametric model in the area of interest is nonlinear, as is true in many fields, there are some advantages to taking a localization of this in the nonparametric approach.

The issue of identification in local regression models is not well explored but some results are known (see Gozalo and Linton 2000). The expansion (4) is clearly crucial for identification. If the function m is continuous but not differentiable, then only a single parameter is identifiable, which corresponds to the first term in (4); additional parameters remain unidentified. It is also necessary that there is a neighbourhood of the estimation point that contains enough observations (this is guaranteed when the marginal density exists and is positive).

Estimation of local regression models can be carried out by localization of the usual estimation criteria adopted for estimation of the corresponding parametric model like maximum likelihood or the method of moments where the localization is carried out by multiplying the contribution of observation i to the sample average objective function by the weight $w_{ni} = K((x - Xi)/h)$, where K is called the kernel and usually satisfies at least $\int K(u)\,du = 1$, while $h = h(n)$ is the bandwidth, a sequence designed to go to zero with sample size. The effect of the weighting factor wni is to emphasize observations close to the point of interest x and to de-emphasize observations far from x, whence the appellation 'localization'.

In the multivariate case, the expansion (4) becomes much more complicated: there are d first order partial derivatives, $d(d - 1)/2$ second order partial derivatives, and so on. With $p = 5$ and $d = 10$ the local parametric model would have over 1,000 parameters, which is too many for practical use. There are many interesting and important cases lying between the two extremes of parametric and fully nonparametric models, where some of the θ_j

vary with only a subset of x. In this case, the local parametric model is imposing exclusion restrictions on the function m and the expansion is reduced. We next give some examples.

A function $m(x)$ is additively separable if

$$m(x) = \sum_{j=1}^{d} m_j(x_j)$$

for some functions m_j. In terms of the framework of the previous section $p = d$ and

$$M(x;\theta) = \sum_{j=1}^{d} M_j(x_j,\theta_j);\theta_j(x) = \theta_j(x_j).$$

The functions $\theta_j(x_j)$ are one-dimensional but of unknown form. This implies that $m(x) = \sum_{j=1}^{d} m_j(x_j)$, where $m_j(x_j) = \theta_j(x_j)$. In this case, each function $\theta_j(x)$ has $d-1$ exclusion restrictions. This is consistent with strong separability as defined in Goldman and Uzawa (1964). A generalization of this is to the so-called generalized additive models where $M(x;\theta) = G\left(\sum_{j=1}^{d} M_j(x_j,\theta_j)\right)$, where $\theta_j(x) = \theta_j(x_j)$, in which G is a known 'link' function, while θ_j are univariate functions as before. For example, G could be the c.d.f. of a random variable like the normal or logit. Linton and Nielsen (1995) discuss estimation of additive models.

In time series one is often interested in the relationship

$$E[y_t|I_{t-1}] = m(I_{t-1}),$$

where the information set $I_{t-1} = \{y_{t-1}, \ldots\}$ includes all past variables, either for estimation or forecasting purposes. This situation is complicated because I_{t-1} contains infinitely many variables and apart from the important class of Markov models m generally depends on all of them. A common assumption here is some kind of mixing condition that guarantees that the effect of y_{t-k} on y_t dies out as $k \to \infty$. For example, an invertible $MA(1)$ process has $m(I_{t-1}) = \sum_{j=1}^{\infty} \theta^{j-1} y_{t-j}$ for some $|\theta| < 1$. A natural generalization of this is the model $m(I_{t-1}) = \sum_{j=1}^{\infty} m_j\left(y_{t-j}\right)$, where m_j is a sequence of functions such that the sum is well defined, that is, $m_j(\cdot)$ must decline in importance as $j \to \infty$. This model is hard to analyse and to estimate. Instead, consider the more restrictive version

$$m(I_{t-1}) = \sum_{j=1}^{\infty} \theta^{j-1} m\left(y_{t-j}\right) \tag{5}$$

for some unknown function $m(\cdot)$ and parameter θ. When $m(y) = y$ this includes the MA(1) process as a special case, but includes many other nonlinear models. By taking a local parametric model $M(y) = a_0(y) + a_1(y)y + a_2(y)y^2$ for m one can nest the GARCH(1, 1) model of Bollerslev (1986). Linton and Mammen (2005) have recently developed a theory of estimation for this class of models.

Another popular approach is the locally stationary models pioneered by Dahlhaus (1997). A locally stationary AR(1) process is $y_t = \rho(t/T)y_{t-1} + \varepsilon_t$, where ε_t is i.i.d. and $\rho(\cdot)$ is a smooth but unknown form. By taking the local parametric model $M(y) = a_0$ one can nest the conventional autoregression, although there are other possibilities. Dahlhaus actually deals with a more general class of linear processes with $y_t = \sum_{j=0}^{\infty} c_j(t/T)\varepsilon_{t-j}$, where $c_j(\cdot)$ are unknown but smooth functions.

See Also

- ▶ Kernel Estimators in Econometrics
- ▶ Non-parametric Structural Models
- ▶ Semiparametric Estimation

Bibliography

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–327.

Charnes, A., W. Cooper, and A. Schinnar. 1976. A theorem on homogeneous functions and extended Cobb–Douglas forms. *Proceedings of the National Academy of Science, USA* 73: 3747–4748.

Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *Annals of Statistics* 25: 1–37.

Delgado, M., and F. Hidalgo. 2000. Nonparametric inference on structural breaks. *Journal of Econometrics* 96: 113–144.
Fan, J., and I. Gijbels. 1996. *Local polynomial modelling and its applications*. London: Chapman and Hall.
Goldman, S., and H. Uzawa. 1964. A note on separability and demand analysis. *Econometrica* 32: 387–398.
Gozalo, P., and O. Linton. 2000. Local nonlinear least squares estimation: Using parametric information nonparametrically. *Journal of Econometrics* 99: 63–106.
Härdle, W., and O. Linton. 1994. Applied nonparametric methods. In *The handbook of econometrics*, vol. 4, ed. D. McFadden and R. Engle. Amsterdam: North-Holland.
Hjort, N., and I. Glad. 1995. Nonparametric density estimation with a parametric start. *Annals of Statistics* 23: 882–904.
Linton, O., and E. Mammen. 2005. Estimating semiparametric ARCH models by kernel smoothing methods. *Econometrica* 73: 771–836.
Linton, O., and J. Nielsen. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82: 93–100.
Rothenberg, T. 1971. Identification in parametric models. *Econometrica* 39: 577–591.

Location of Economic Activity

M. J. Beckmann

Location theory develops principles for determining where the various economic activites take place. It was focused on the location of types of land use in agriculture (von Thünen 1826) before addressing the emerging problems of the locational concentration of 'heavy' and related industries (Launhardt 1882; Weber 1909) and the shifts of industrial production locations induced by the railroad system and the growth of international trade. The first systematic treatment of industrial location theory was due to Alfred Weber (1909). Location theory was treated as a branch of price theory by Engländer (1924) and Predöhl (1925). This development culminated in Palander's theory of imperfect competition in spatial markets (1935). A. Lösch (1940) expanded the scope of location theory to include the location of services in 'central places', a concept due to W. Christaller (1933), of trade flows and of transportation networks. The systematic study of the location of consumers (households or residential land use) had to await the development of the 'new' urban economics (Alonso, Mills, Beckmann). A macroeconomic perspective of locational problems was provided by regional economics – the study of regional differences in population, output, income, capital and growth (Isard 1960). The development of methods for the actual calculation of the optimal location for single facilities or systems of facilities (warehouses, assembly plants, Disneyland, etc.) was undertaken independently of economics by operations researchers, utilizing such new tools as linear and integer programming which are appropriate when the search for optimum locations is restricted to a finite set of points on a grid or in a network (nodes).

The location problem may be studied in an international context (as specialization in international trade) or in an urban context (as urban economics), but it is usually treated for a national economy.

The following exposition uses the framework of price theory to integrate as far as possible location theory into general economic theory. In this view location theory becomes spatial microeconomics. (Macroeconomic aspects are treated in *regional science*.)

The explosive growth of the location theoretic literature forces us to present only the basic problems for each category in some detail and to give brief summaries of further developments.

Spatial Supply and Demand

The analysis of spatial supply and demand is the theory of location in the short run. It revolves around the question: Which demand is satisfied through local production, which through imports, and which will not be satisfied at all?

It turns out that some products are made in almost all locations for local consumption only while some are produced in specialized locations and shipped to larger markets. Beyond critical distances from supply points, some goods are not available at all. Which of these events happens depends on: size of local demand, transportability

of product, availability of resources, cost of resources, transportability of resources.

The modern approach takes a closer look at the interaction of availability and transportability through *linear programming*. This is based on the following simple spatial-allocation problem known as the transportation problem of Hitchcock (1941), Kantorovich (1942) and Koopmans (1949). It is in fact a short-run location problem in which one determines how the production of a commodity is spatially distributed within given capacity limits. Given are the quantities q_j of a commodity demanded in a set of locations j, $j = 1,\ldots, n$, and the capacities c_i of plants i for producing this commodity, $i = 1,\ldots, m$. Location enters through the matrix of transportation costs t_y for transporting the commodity from production locations i to consumption locations j. Feasibility requires that total capacity be not less than total demand $\Sigma_j q_j \leqslant \Sigma_i c_i$. Given this condition, a solution always exists which minimizes total transportation costs. It can be characterized by means of efficiency prices, which may be interpreted as competitive market prices. Shipments from i to j should be made within the limits imposed by capacity and demand quantities if and only if the price difference equals transportation cost. At points where capacity is underutilized (excess supply) the price of the product is zero.

When production costs differ among plant locations, the object of minimization (by the market or under planning) is the sum of production and transportation costs. The efficiency prices now equal local production cost in places where production takes place at levels below capacity, and equals or exceeds production cost where capacity is fully utilized.

The distinction between production and consumption locations is arbitrary and may be dropped when consumption and production may take place everywhere. The efficiency conditions in terms of the efficiency prices then imply that cross-hauling is always inefficient. Moreover, there exist solutions such that locations which import do not export, and vice versa. The last property is no longer true when transportation is restricted to an existing network. In that case there may be transshipment points.

So far the location problem has been treated as a supply problem only, since demand was assumed given. While this is in the spirit of traditional location theory, it is contrary to neoclassical economics since it neglects the effects of price on market demand.

If demand at each location is considered a function of the commodity price at that location only (interlocal effects on income ignored), the resulting spatial market-equilibrium problem may be obtained as solution to a non-linear programming problem: the maximum of consumers' surplus in demand locations minus transportation and production costs and subject to flow constraints (Samuelson 1952). If demand is bounded and the local price exceeds this bound, then the commodity will not be demanded there and not shipped there. Conversely, if price is below production costs in a plant location, the commodity will not be produced there. When all excess supply curves are equal, then production is for local consumption only.

When supply is concentrated and demand is dispersed, it is natural to consider the sets of locations to which a given supplier ships his product. When transportation costs increase with distance, this set tends to fill a contiguous area, the *market area* surrounding the supplier. The entire region is then divided into mutually exclusive and non-overlapping market areas. When transportation costs fall, low-cost suppliers will expand their market areas into those of high-cost suppliers will (assuming perfect competition). In extreme cases only one supplier survives, an event which is made more likely when production is subject to increasing returns to scale.

When demand is concentrated and supply dispersed, as in labour markets, demand points are surrounded by so-called supply areas. An example is the von Thünen model of a central city and its agricultural hinterland.

Spatial Pricing and Output

The linear and non-linear programming models determine the spatial distribution of output and competitive market prices in the short run. Prices are subject to the same laws in the long run.

When producers are isolated at centres of market areas, they enjoy spatial monopoly even when other firms exist, provided their potential market areas do not overlap.

Three basic price strategies exist: mill pricing or f.o.b. when customers are charged full transportation cost; uniform pricing or c.i.f., when firms charge the same inclusive price to all customers but may refuse to supply them; and perfectly discriminatory pricing, when the price charged depends on distance but does not reflect the full transportation cost. Suppose that all customers have the same linear demand curve but may be located at different distances. Then mill pricing and uniform pricing are equally profitable, while discriminatory pricing (which turns out to be a simple average of the two) is more profitable. For a given market radius, profit-maximizing output is the same under all three price strategies.

Mill pricing is best, discriminatory pricing intermediate and uniform pricing worst in regard to the following measures of social efficiency: total transportation cost; aggregate consumer expenditure; average price paid per commodity unit; consumers' surplus and social surplus, that is the sum of consumers' surplus and profit (see Beckmann 1976).

When the market radius may also be chosen, discriminatory pricing can be best since it extends the market farther (Holahan 1975). The socially optimal pricing scheme is mill pricing at marginal cost, but with constant or decreasing marginal cost, firms would not recover their fixed cost. Ramsey pricing (which maximizes the social surplus subject to cost recovery by firms) turns out to be a combination of mill and discriminatory pricing, with the latter component more prominent the higher the fixed-cost level to be met.

Results for non-linear demand functions are few. For example, when demand functions are convex (concave) any mill price is more (less) profitable than a uniform price that exceeds the mill price by the amount of average transportation cost, and hence the optimal mill price (uniform price) must be even more profitable (Stevens and Rydell 1966). Also the following proposition holds: average transportation cost per unit sold is lower under any system of mill pricing than under any system of uniform pricing (Beckmann 1985a). This remains true even when the mill or uniform prices are not the profit-maximizing prices. Comparison of other economic variables (such as output and welfare) have not been made and would require additional assumptions.

Spatial duopoly: the Hotelling problem. Let the buyers of a commodity be:

> uniformly distributed along a line of length l (which may be Main Street in a town or a transcontinental railroad). At distances a and b, respectively, from the two ends of the line are the two places of business A and B. Each buyer transports his purchase home at a cost t per unit distance. . . . Suppose that the cost of production is zero and that unit quantity of the commodity is consumed in each unit of time in each unit length of line. No customer has any preference for either seller except on the ground of price plus transportation cost. . . . The point of division between the regions served by the two entrepreneurs is determined by the condition that at this place it is a matter of indifference whether one buys from A or from B. . . . Each competitor adjusts his price so that with the existing value of the other price his own profit will be a maximum (Hotelling, pp. 107–9).

This is a non-cooperative two-person game whose strategies are the mill prices $p1$, $p2$ of the duopolists. This game has no Nash equilibrium in pure strategies, contrary to Hotelling's claim (d'Aspremont et al. 1979). When quantity demanded depends on price, an equilibrium exists, provided the firms are spaced so far apart that markets do not overlap when each firm charges its monopoly price. The same is true for uniform pricing. This invalidates Hotelling's (1929) belief that spatial differentiation yields stability.

Existence is not restored simply by eliminating discontinuities in the demand functions. However, under *discriminatory pricing* and inelastic demand, an equilibrium exists. For each location compare the lowest price at which firm A and B can supply the product. The higher price minus ϵ is the equilibrium price, and the lower-cost firm is the supplier. This implies that in the contested market areas prices actually rise with decreasing distance for the actual supplier (Hoover 1937; Beckmann 1968). For spatial duopoly, examples may be constructed such that each consumer is better off under discriminatory

pricing than under mill pricing. While price discrimination creates stability, it may (in special cases) also increase social welfare as measured by the consumers and producers surplus.

Locational Choice

The locations of production activities are not predetermined but are subject to economic choice.

In general, availability of resources, location of population as a source of labour and as potential markets, soil, climate and technical conditions rule out many locations for any particular economic activity. What remains is a set of feasible locations among which an economic choice is to be made. In neoclassical economic theory these choices are seen as attempts to maximize profits.

Depending on how location influences revenues and costs, four situations arise that have been treated at various depth in the location theoretic literature (Lösch 1940; Hoover 1948; Isard 1956; Greenhut 1956; Beckmann 1968).

Input (resource or labour) orientation means that either input availability or input prices and hence costs vary with location, and that the firm can sell all its output at the same price everywhere. Its optimal location is then that of minimal production costs. If production requires a localized resource as input and all other input prices are uniform, then the production activity is drawn to the resource location – for all other locations would require costly transportation of the resource. This remains true even when product prices are not strictly uniform but differ only slightly among locations.

Footloose implies that both product prices and factor prices are uniform. This means in particular that either no materials are used as inputs (as in the case of services) or that the material inputs are available everywhere at constant prices. They are then called ubiquities.

Market orientation means that production costs are equal at all locations, but the product requires costly transportation because, due to economies of scale in production, consumers are more dispersed than producers, each firm then serving a separate market area. Determining the size of these is the basic problem of the following section. Market orientation can also result when resources are localized but, after some initial processing are easier to transport than finished products (e.g. wood, metal ingots).

It has been argued that market orientation is replacing resource orientation for many economic activities. Lösch (1940) has pointed out that nationwide bargaining for wages has tended to eliminate interregional wage differences, and thus any labour orientation that may have existed.

The general case in the south-east corner of Table 1 includes one special situation that has been the focus of classical location theory: suppose that the locations of markets for product and of resources are all given. Moreover, let product quantities at markets and factor quantities at resource deposits be given and fixed, and let wages be independent of location. In that case, profit maximization reduces to cost minimization, and cost minimization in turn reduces to the minimization of transportation costs. The classical location analysis of Launhardt and (subsequently and independently) of A. Weber considers this case.

The optimal location is uniquely determined by the following condition. Draw connecting lines from the unknown location L to the input and market locations (in the simplest case there are just three) – A B C – and consider them as lines of force. The forces are the weights to be hauled w_A, w_B, 1. The necessary and sufficient condition for an optimal location at an interior point is that the three forces be in equilibrium.

Location of Economic Activity, Table 1

		Revenue	
		Independent of location	Dependent on location
Cost	Independent of location	Footloose activity	Market orientation
	Dependent on location	Resource or labour orientation	no single orientation

The problem of minimizing transportation cost may be solved by an analogue device credited to Varignon. Let three rollers be placed on the periphery of a circular device, in locations corresponding to the three given ones. Three pieces of string are tied by a single knot, and weights are attached to the ends of the strings in proportion to the weights that are moved in the problem. When the strings are put over the corresponding rollers the knot will be pulled into that point where the potential energy of the system is minimal – that is, where the weights hang down as far as possible – and this is equivalent to the minimization of the distances weighted by the respective forces.

If one force is large enough, it will pull the knot right up to one roller. This is always the case when one weight equals or exceeds the sum of the other weights. Thus production which is weight-increasing – that is, such that output weighs more than all inputs taken together – should always take place at the point of consumption; it should be market-oriented.

The principle of the equilibrium of weights considered as forces still applies, when there are more than two inputs and more markets, provided the quantities to be shipped to the various markets are fixed. When the assumption of fixed coefficients is dropped, the proportions of inputs will vary with location, and the Weberian equilibrium conditions must be supplemented by the usual conditions equating marginal rates of substitution to factor-price ratios (Predöhl 1925; Isard 1956; Moses 1958). When demand is price dependent, the optimal location for the firm is no longer at the point of minimum cost but of maximum profit.

Suppose the choice of locations is restricted to points on a given transportation network. The optimum location is once more a point where all forces that represent weights to be moved are in equilibrium. There may be many points at which this condition is satisfied, but they are all found to be nodes (Hakimi 1964).

As a special case, let the network consist of a straight line on which resources and market locations are marked off as discrete points. Then the optimal location is determined by the so-called *principle of the median*: after a small displacement to the left (right), the forces pulling to the right (left) must exceed those pulling to the left (right). In particular, if all points are markets (and the material to be processed is ubiquitous), then the principle of the median yields as solution the median point of the demand distribution.

If the network consists of a rectangular road grid with east-west and north-south roads intersecting at right angles, then the principle of the median applies separately in each of the two principal directions.

Spatial Duopoly

The negative results concerning the existence of an equilibrium for the Hotelling problem carry over to the case when the two sellers may choose both price and location. Suppose, however, that prices are fixed and identical for the two firms (Lerner and Singer 1937; Beckmann 1968; Eaton and Lipsey 1976) so that they compete on location only. (This model has been applied to the choice of political platforms in party competition; Smithies 1941). In this case, while welfare maximization would require location at the first and third quartiles, there is an equilibrium with both sellers located together at the median point.

Spatial Equilibrium of an Industry

In the long run the locational pattern of firms in an industry is governed by free entry and exit: not only the location but also the number of firms is variable. This implies an equilibrium in which profits are zero (or as close to zero as consistent with integer numbers of firms in a bounded region). In a two-dimension setting both the *size* and the *shape* of equilibrium market areas must be determined, and how they depend on the type of competition and on the demand and cost parameters of the problem. The size is usually discussed in a one-dimensional context.

The standard model is this. Firms produce a homogeneous product under positive fixed costs, F, and constant marginal costs c, and the transport rate, t is constant. Consumers are uniformly

distributed along an unbounded linear market (or, equivalently, along a circular market) with a density a; they have identical (downward-sloping) demand functions, $f[p(x)]$. Each consumer buys from the firm offering the product at the lowest full price. The market radius R of a firm whose two neighbours are at equal distance D charging identical prices $\overline{p}$ is given by

$$R = \frac{\overline{p} - p + tD}{2t}.$$

The profit-maximizing price depends on the reaction function $\overline{p}(p)$ of the other firms. Two extreme cases are considered: Bertrand response $d\overline{p}/dp = 0$ and Löschian response $d\overline{p}/dp = 1$. For rectangular demand curves (a unit quantity is bought at a reservation price of unity), the equilibrium price and spacing can be written in closed form:

$$\text{Bertrand}: D^* = \sqrt{\frac{F}{at}} \qquad p^* = c + tD^*$$

$$\text{Lösch:} \quad D^* = \frac{1-c}{t} - \sqrt{\left(\frac{1-c}{t}\right)^2 - \frac{F}{at}},$$

$$p^* = \frac{1}{2}\left(1 + c + \sqrt{(1-c)^2 - \frac{Ft}{a}}\right).$$

The more profitable Löschian strategy leads to smaller market areas and higher mill prices but also to a larger consumers' surplus. The industry can survive only when fixed cost is small enough

$$F \leqq \mathrm{a}\frac{(1-\mathrm{c})^2}{2t}$$

The comparative static properties of the Löschian solution are these: the market radius increases with all costs and decreases with consumer density and reservation price. The mill price behaves in just the opposite way, which may appear counter-intuitive.

Neither Bertrand nor Löschian pricing with free entry results in a social optimum, and mill pricing at marginal cost will not recover fixed costs, but uniform pricing at marginal cost (of production and delivery to the most distant customer) does and constitutes a second-best solution.

A question much debated concerns the shape of market areas in long-run equilibrium when firms will adjust their locations perfectly and profits are driven exactly the zero.

When possible market areas are restricted to the three regular polygons: equilateral triangles, squares and hexagons, then hexagons represent the most profitable shape since, for many cost and demand functions, profits in a given area are a decreasing function of the average distance to customers, and for a given area this average distance is smallest in a hexagon (Lösch 1940).

But average distance is smaller still in a circle. When fixed cost are larger than the gross profits that can be achieved in a hexagon but smaller than those attainable in a circle, the question arises as to the equilibrium shape of the market area (Mills and Lav 1964).

Firms operating at first in the most profitable circular markets, being squeezed by new entrants, find their market areas reduced to rounded hexagons. Only when fixed costs are small enough does the squeeze continue until hexagons emerge. Firms do not sell in those points of the plane where demand is zero, because mill price plus transport costs exceeds the price intercept $f^{-1}(0)$ (Beckmann 1971; Mulligan 1981). As limiting cases, we have a complete hexagon, the traditional Löschian market area, and at the other extreme a complete circle, the monopoly market area.

When the entry process is treated explicitly as a sequential process taking place over time, this entails an asymmetry in the choices open to the existing firms and to the entrants. Whereas the former have to stick to their location, the latter are free to choose where to set up their plant. This has several important implications regarding the properties of the long-run equilibrium. First, free-entry is consistent with the persistence of pure profits (Eaton and Lipsey 1976) since the market that a newcomer can capture may be too small to recover his set-up costs. Second, the long-run equilibrium is not unique in that it depends on the initial conditions and on the dynamics of

entry. In particular, the regular spacing of firms turns out to be destroyed in many cases so that firms can enjoy market areas of different sizes and shapes at equilibrium. Third, entry can be deterred (hence profits can be increased) if firms are sophisticated enough to take full advantage of their position in the order of entry (see Hay 1976). In so doing, a firm locates so as to secure the most profitable market for itself in the long run. This implies that the firms are fewer than under myopic behaviour. In the same vein, a monopolist with multiple plants can pre-empt the entire profits, subject to the constraint that no additional firms can earn a profit after entry (Prescott and Visscher 1977).

Operations researchers have developed a literature on *planning solutions* when choice is restricted to a finite set of locations both for a single plant and for the multiple-plants location problem (also known as the location-allocation problem). (For a survey, see Krarup and Pruzan 1985.)

Formally, this is a programming problem combining both zero-one variables (assigning plants to locations), and continuous variables describing commodity flows (Stollsteimer 1963; Manne 1964). Numerical methods such as branch and bound and linear programming relaxation have been successfully applied to problems whose size is not too large. Heuristic methods, based on experimentation through trial and error are applicable to much larger problems (Erlenkotter 1978). Although they do not guarantee an optimum solution, the achieved total cost of the heuristic solution approximates the absolute minimum reasonable well in most cases.

Spatial-Resource Use

The classical analysis of the allocation of land to competing land-using activities is that of von Thünen:

> consider a very large town in the center of a fertile plain. . . . The soil of the plain is assumed to be of uniform fertility which allows cultivation everywhere. At a great distance the plain ends in an uncultivated wilderness, by which this state is absolutely cut off from the rest of the world. The question is: How . . . will distance from the city affect agriculture methods when these are chosen in the optimal manner? (von Thünen 1826, pp. 11–12).

If transportation cost is proportional to weight and distance, the answer is as follows. The various products are grown in zones (rings) of exclusive specialized land-use. The zones are arranged in the order of decreasing weights produced per hectare, and their width is determined by the quantities demanded. Thus emerges the typical sequence of truck gardening, milk production, cereals, grazing areas and forests.

This is true not only when (as assumed by von Thünen) input coefficients are constant but when returns to scale are constant or increasing. A graphical analysis shows that the rent bids by the various products for land use are a linear decreasing function of distance, and that a product with steeper descent can outbid another only at points closer to the centre. The upper contour of these bid curves representing the successful rent bids – i.e. the emerging market rent's – is downward-sloping and convex and falls to zero at the distance beyond which production for the market is unprofitable.

The von Thünen analysis may be applied within the city as a model of residential land use, yielding a sequence of, for example, high-rise apartment houses, multi-storey row houses, free-standing two-storey houses and single-storey houses, and ranch-type homes such that population density decreases with distance from the central business district (Beckmann 1968).

The von Thünen model of land use may be extended to multiple centres in an obvious way. A more challenging situation is that where the demands are also dispersed in the manner of a continuous density. An extended spatial market of this type may be described (analogously to physics) by a field of commodity flows with demands acting as 'sinks' and supplies acting as 'sources'. This flow field then defines a 'potential function' which represents the local prices. The direction of flow is the gradient direction of this potential function. Two important economic conclusions result: the flow field arranges itself as a

set of market and supply areas, although the centres of these no longer represent isolated cities (Beckmann 1952; Beckmann and Puu 1985). Secondly, land use is once more specialized. This specialization does not require fixed coefficients of production but is a consequence of any linear homogeneous production technology. The main theme of spatial-resource theory is thus specialization of land use and the importance of market access. Agriculture emerges as a market-oriented industry.

General Spatial Equilibrium

This is usually considered for particular systems such as a single city (and its internal structure), or a city and its hinterland (a reinterpreted von Thünen model) or a system of cities in a homogeneous spatial setting, a *central place system*. All these systems show as a new phenomenon the spatial separation of unrelated, and the clustering of related, economic activities. Separation results from the bidding for land use. A clustering of plants from different industries is a common phenomenon. Clustering occurs under the following circumstances:

1. The products are unrelated but the spacing of firms in long-run equilibrium is approximately the same. If customers were truly dispersed at uniform density, then industries have no incentive to cluster; on the contrary, competition for land and labour would draw them apart. But the presence of one industry creates a local market for the product of the other.
2. The labour markets of the two industries may be complementary: one industry using male labour (mental fabricating) while the other employs female labour (custom jewellery, lacemaking).
3. The output of one industry is used exclusively as an input by the other industry.
4. The two industries use a common resource which is less transportable than the output.
5. There are economies of joint location arising from the use of special facilities, such as transportation, marketing, training and other auxiliary functions.
6. There are advantages of sharing the social infrastructure. These various cost savings are known as *agglomeration economies* (Weber 1909).

Another approach to the positioning of related economic activities starts out by assuming that there is a given number of discrete locational 'sites' or land lots, and a finite number of economic activities or 'plants' competing for these locations. This is the so-called *assignment problem*. Whether a market exists that can assign plants to locations through a mechanism of competitive prices depends on the presence or absence of 'linkages' between the plants seeking locations.

While linkages may create externalities that prevent a market equilibrium, the spatial structure of agglomeration in cities may be more orderly: clusters composed of interacting plants with close support links may locate apart from other such clusters, when linkages between different but adjacent clusters are weak. Or there may be a mix of plants of widely varying degrees of interaction provided transportation cost is insensitive to distances if they are short enough.

Central Place Model

Suppose next that market areas of an industry A are not equal to but are a small multiple of those of industry B. Then agglomeration economies will cause every A location to contain also a B plant, but some B plants are by themselves. If this is true for any pair of industries, the result will be clusters of various composition.

Suppose, however, that market areas (in terms of population) of plants in all marketoriented industries are all approximately equal to a, ma, m^2a, . . ., m^ra. Then a is the basic market size, m a nesting factor and r the rank of a market and of the centre serving that market. Such a hierarchy of market areas and their centres is called a Central Place system (Christaller 1933; Lösch 1940). Each market-oriented industry is then found in a centre of a characteristic rank and all centres of higher ranks. The location problem for market-oriented industries reduces to finding the rank of a

market where a given economic activity can survive. Higher-order centres will in general have more than one firm of that industry.

These centres 'export' their characteristic product to lower-order centres in their market area. Lowest-order centres serve an agricultural hinterland. There is no trade among centres of equal rank. The national centre exports products and services of the highest order to the entire nation as a region. Of course, this service may not reach all potential customers because 'interaction decreases with distance'.

Location theory has been called the economics of distance. Without transportation cost, there would be no location problem. Indeed, transportation cost and economies of scale are the sole determinants of the location pattern in a homogeneous region. When resources are localized, or impediments to transportation exist, a heterogeneous system of economic activities using localized resources (energy, mineral resources, climate) is superimposed on an otherwise homogeneous central place system. *Economic geography* is the systematic study of the location patterns that result from the actual distribution of resources and the geography of natural transportation barriers and channels.

See Also

- ▶ Central Place Theory
- ▶ Christaller, Walter (1894–1975)
- ▶ Gravity Models
- ▶ Lösch, August (1906–1945)
- ▶ Monocentric Models in Urban Economics
- ▶ Spatial Economics
- ▶ Thünen, Johann Heinrich von (1783–1850)
- ▶ Tiebout Hypothesis
- ▶ Urban Economics
- ▶ Weber, Alfred (1868–1958)

Bibliography

Alonso, W. 1967. A reformulation of classical location theory and its relation to rent theory. *Papers of the Regional Science Association* 19: 23–44.

Beckmann, M.J. 1952. A continuous model of transportation. *Econometrica* 20: 643–660.

Beckmann, M.J. 1968. *Location Theory*. New York: Random House.

Beckmann, M.J. 1969. On the distribution of urban rent and density. *Journal of Economic Theory* 1: 60–67.

Beckmann, M.J. 1971. Equilibrium versus optimum: Spacing of firms and patterns of market areas. *Northeast Regional Science Review* 1: 1–20.

Beckmann, M.J. 1976. Spatial price policies revisited. *Bell Journal of Economics* 7: 619–630.

Beckmann, M.J. 1985a. Spatial price policy and the demand for transportation. *Journal of Regional Science* 25(3): 367–371.

Beckmann, M.J. 1985b. A model of perfect competition in spatial markets. *International Review of Economics and Business* 32(5): 413–419.

Beckmann, M.J. and Marschak, T. 1955. An activity analysis approach to location theory. In: *Proceedings of the second symposium in linear programming*. Washington, DC: National Bureau of Standards and Directorate of Management Analyses.

Beckmann, M.J., and T. Puu. 1985. *Spatial economics: Density, potential and flow*. Amsterdam: North-Holland.

Benson, B.L. 1980. Löschian competition under alternative demand conditions. *American Economic Review* 70: 1098–1105.

Capozza, D.R., and R. Van Order. 1978. A generalized model of spatial competition. *American Economic Review* 68: 896–908.

Christaller, W. 1933. *Die zentralen Orte in Suddeutschland*. Jena: Gustav Fisher. English trans. by C.W. Baskin as *Central places in Southern Germany*. Englewood Cliffs: Prentice-Hall, 1966.

D'Aspremont, C., J. Gabszewicz, and J. Thisse. 1979. On hotelling's 'stability in competition'. *Econometrica* 47(5): 1145–1150.

Eaton, B.C., and R.G. Lipsey. 1976. The non-uniqueness of equilibrium in the Löschian model. *American Economic Review* 66(1): 71–93.

Engländer, O. 1924. *Theorie des Güterverkehrs und der Frachtsätze*. Jena: G. Fischer.

Erlenkotter, D. 1978. A dual-based procedure for uncapacitated facility location. *Operations Research* 16: 992–1009.

Greenhut, M.L. 1956. *Plant location in theory and practice*. Chapel Hill: University of North Carolina Press.

Greenhut, M.L. 1970. *A theory of the firm in economic space*. Austin: Lone Star.

Hakimi, S.L. 1964. Optimum location of switching centers and the absolute centers and medians of a graph. *Operations Research* 12: 450–459.

Hanjoul, P., and J.-F. Thisse. 1984. The location of a firm on a network. In *Applied decision analysis and economic behaviour*, ed. A.J. Hughes Hallet. The Hague: Martinus Nijhoff.

Hay, D.A. 1976. Sequential entry and entry-deterring strategies. *Oxford Economic Papers* 28: 240–257.

Hitchcock, F. 1941. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics* 20: 224–230.

Holahan, W.L. 1975. The welfare effects of spatial price discrimination. *American Economic Review* 65: 498–503.
Holahan, W.L., and R.E. Schuler. 1981. The welfare effects of market shapes in the Lüschian Location Model: squares vs. hexagons. *American Economic Review* 71: 738–746.
Hoover, E.M. 1937. Spatial price discrimination. *Review of Economic Studies* 4: 182–191.
Hoover, E.M. 1948. *The location of economic activity*. New York: McGraw-Hill.
Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.
Hsu, S. 1983. Pricing in an urban spatial monopoly: A general analysis. *Journal of Regional Science* 23: 165–175.
Isard, W. 1956. *Location and space-economy*. New York: Wiley.
Isard, W. 1958. Interregional linear programming. *Journal of Regional Science* 1: 1–59.
Isard, W. 1960. *Methods of regional analysis: An introduction to regional science*. New York: Wiley and Technology Press.
Kantorovich, L. 1942. On the translocation of masses. *Doklady Akademii Nauk CCCP* 37: 199–201. Reprinted in *Management science* 5, October 1958: 1–4.
Koopmans, T.C. 1949. Optimum utilization of the transportation system. *Econometrica* 17(Supplement): 136–146.
Koopmans, T.C., and M.J. Beckmann. 1957. Assignment problems and the location of economic activities. *Econometrica* 25: 53–76.
Krarup, J., and P.M. Pruzan. 1985. Ingredients of locational analysis. In *Discrete location theory*, ed. R.L. Francis and P. Mirchandani. New York: Wiley and Interscience.
Launhardt, W. 1882. Die Bestimmung des zweckmässigsten Standortes einer gewerblichen Anlage. *Zeitschrift des Vereins Deutscher Ingenieure* 26: 106–115.
Lerner, A., and H.W. Singer. 1937. Some notes on duopoly and spatial competition. *Journal of Political Economy* 45: 145–186.
Lösch, A. 1940. *Die räumliche Ordnung der Wirtschaft*. Jena: Gustav Fisher. Trans. as *The economics of location*. New Haven.: Yale University Press, 1954.
Manne, A.S. 1964. Plant location under economies of scale: decentralization and computation. *Management Science* 11: 213–235.
Mills, E.S. 1972. *Studies in the structure of the urban economy*. Baltimore: Johns Hopkins Press.
Mills, E.S., and M.R. Lav. 1964. A model of market areas with free entry. *Journal of Political Economy* 72: 278–288.
Moses, L. 1958. Location and the theory of production. *Quarterly Journal of Economics* 72: 259–272.
Mulligan, G.F. 1981. Lösch's single-good equilibrium. *Annals of the Association of American Geographers* 71: 84–94.
Ohta, H. 1981. The price effects of spatial competition. *Review of Economic Studies* 48: 317–325.
Paelinck, J.H.P., J.P. Ancott, and J.H. Kinpur. 1983. *Formal spatial analysis*. Aldershot: Gower.
Palander, T. 1935. *Beiträge zur Standortstheorie*. Uppsala: Almqvist & Wiksell.
Ponsard, C. 1983. *A history of spatial economic theory*. Berlin: Springer-Verlag.
Predöhl, A. 1925. Das Standortsproblem in der Wirtschaftstheorie. *Weltwirtschaftliches Archiv* 21: 294–331.
Prescott, E.C., and M. Visscher. 1977. Sequential location among firms with foresight. *Bell Journal of Economics* 8: 378–393.
Puu, T. 1979. *The allocation of road capital in two-dimensional space*. Amsterdam: North-Holland.
Samuelson, P.A. 1952. Spatial price equilibrium and linear programming. *American Economic Review* 42: 283–303.
Singer, H.W. 1937. A note on spatial price discrimination. *Review of Economic Studies* 5: 75–77.
Smithies, A. 1941. Monopolistic price policy in a spatial market. *Econometrica* 9: 63–73.
Stevens, B.H., and C.P. Rydell. 1966. Spatial demand theory and monopoly price policy. *Papers of the Regional Science Association* 17: 195–204.
Stollsteimer, J.F. 1963. A working model for plant numbers and location. *Journal of Farm Economics* 43: 631–645.
von Thünen, J.H. 1826. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Stuttgart: Gustav Fischer, 1966.
Weber, A. 1909. *Ueber den Standort der Industrien*. Tubingen: J.C.B. Mohr. Trans. C.J. Friedrich as *Theory of the location of industries*. Chicago: University of Chicago Press, 1929.

Location Theory

Jacques-François Thisse

Abstract

Location theory deals with *what* is *where*. 'What' refers to any possible type of economic activity involving stores, dwellings, plants, offices, or public facilities. 'Where' refers to areas such as regions, cities, political jurisdictions, or custom unions. The objective of location theory is to explain why particular economic activities choose to establish themselves in particular places. Here we focus on spatial competition theory between firms,

where locations are subject to attracting and repelling forces. We then extend this framework in order to account for the residential choices made by consumers.

Keywords

Agglomeration forces; Clusters; Collusion; Dispersion forces; General equilibrium; Globalization; Hotelling, H; Land use; Local labour markets; Location theory; Nash equilibrium; New economic geography; Oligopolistic competition; Partial equilibrium; Principle of Differentiation; Product differentiation; Spatial competition; Stigler, G; Transportation costs; Urban economics

JEL Classifications

R10

From a historical perspective, location theory has been at both the centre and the periphery of economic theory. It has been at the centre to the extent that it has followed the tradition taking its roots in Hotelling's classical paper 'Stability in Competition' (1929) and has used the spatial framework as a metaphor to explain issues involving heterogeneity and diversity across agents (Rosen 2002). Examples include the supply of differentiated products, electoral competition between political parties, the matching process on the labour market, competition between communities to attract residents or firms, and the number and size of jurisdictions. Location theory has been at the periphery to the extent that space has not been a major concern for most economists. Indeed, it is rare to find a principles textbook in which location issues are covered, let alone mentioned. This is despite their obvious importance for the way actual markets function, as shown, for instance, by the debate raging in many industrialized countries about the consequences of globalization for the location of jobs.

The theory of optimal location for a firm has long been dominated by the minisum model in which the firm aims at minimizing its total transportation costs (Weber 1909). Formally, this is achieved by minimizing the weighted sum of distances to a finite number of points, which represent input and output markets. When the length of the shortest path connecting any two points of a transportation network measures the distance between these points, the firm's optimal location is an input/output market, or a node of the network, or both (Hurter and Martinich 1988). Hence, *the locational choice of a firm is either sluggish or catastrophic*. Another interesting feature of that model is that the firm's optimal location is the outcome of the interplay of a system of forces pulling the firm in different directions. When several competing firms are to be located, the system of forces becomes richer in that it involves what are called 'agglomeration' and 'dispersion' forces.

Spatial Competition Between Firms

To see how such a system of forces works, we consider the framework developed by Hotelling (1929). The market of a homogeneous good is made up of consumers who request one unit of the good. Because any single consumer is negligible to firms, Hotelling assumes that consumers are continuously distributed along a linear and bounded segment: think of Main Street. For simplicity, consumers are also supposed to be uniformly distributed along the linear segment. Two stores, aiming to maximize their respective profits, seek a location along the same segment. Because they are dispersed across locations, consumers differ in their access to the same store. In such a context, firms anticipate correctly that each consumer will buy from the store posting the lower full price, namely, the price at the firm's gate, called 'mill price', augmented by the travel costs that consumers must bear to go to the store they patronize. Accordingly, once they are located firms have some monopoly power over the consumers located in their vicinity, which enables them to choose their price. Of course, this choice is restricted by the possibility that consumers have to supply themselves from the competing firm. Note that any firm is supposed to have a single location – that is, an *address* – because increasing returns and indivisibilities do not allow it to run a

large number of outlets dispersed along Main Street without incurring major losses (Koopmans 1957).

Since each firm is aware that its price choice affects the consumer segment supplied by its rival, *spatial competition is inherently strategic*. This is one of the main innovations introduced by Hotelling, who uses a two-stage game to model the process of spatial competition. In the first stage, stores choose their location non- cooperatively; in the second, these locations being publicly observed, firms select their selling price. The use of a sequential procedure means that firms anticipate the consequences of their locational choices on their subsequent choices of prices, thus imparting to the model an implicit dynamic structure. The game is solved by backward induction. For an arbitrary pair of locations, Hotelling starts by solving the price subgame corresponding to the second stage. The resulting equilibrium prices are introduced into the profit functions, which then depend only upon the locations chosen by the firms. These functions stand for the payoffs that firms will maximize during the first stage of the game. Such an approach anticipates by several decades the concept of subgame perfect Nash equilibrium introduced by Selten (1965).

Whereas the individual purchase decision is discontinuous – a consumer buying only from one firm – Hotelling finds it reasonable to suppose that firms' aggregated demands are continuous with respect to prices. Supposing that each consumer is negligible solves the apparent contradiction between discontinuity at the individual level and continuity at the aggregated level. In other words, when consumers are continuously distributed across locations aggregated demands are 'often' continuous. The hypothesis of the continuum that had been popularized much later by Aumann (1964) is found here to represent the idea that competitive agents have a negligible impact on the market outcome. However, Hotelling considers a richer setting involving both 'dwarfs' – consumers – whose behaviour is competitive and 'giants' – firms – whose behaviour is strategic because they can manipulate the market outcome.

Hotelling's claim was that the process of spatial competition leads firms to agglomerate at the market centre. If true, this provides us with a rationale for the observed spatial concentration of firms selling similar goods (such as restaurants, movie theatres, or fashion clothes shops). But Hotelling's analysis is undermined by a mistake that invalidates his main conclusion: when firms are sufficiently close, the corresponding subgame does not have a Nash equilibrium in pure strategies, so that the payoffs used by Hotelling in the first stage are wrong (d'Aspremont et al. 1979). This negative conclusion has led d'Aspremont et al. to slightly modify the Hotelling setting by assuming that the travel costs borne by consumers are quadratic in the distance covered, instead of being linear as in Hotelling. This new assumption captures the idea that the marginal cost of time increases with the length of the trip to the store. In this modified version, d'Aspremont et al. show that any price subgame has one and only one Nash equilibrium in pure strategies. Plugging these prices into the profit functions, they show that firms choose to set up at the two extremities of the linear segment. Firms do so because this allows them to relax price competition and to restore their profit margins. Indeed, when prices are fixed and equal the quest for customer proximity – or, equivalently, for a larger market area – leads the two firms to agglomerate at the market centre. The tendency for firms to choose distinct locations or products has been confirmed by many works, and has led Tirole (1988) to call it the 'Principle of Differentiation'.

Consequently, *price competition is a dispersion force*, whereas *the market area effect is an agglomeration force*. What the Principle of Differentiation tells us is that the dispersion force always dominates the agglomeration force, at least when firms sell a homogeneous product and compete in price. Hence, the Hotelling setting is to be enriched if we want to be able to understand why firms selling similar products often form spatial clusters. This has been accomplished by following two different research strategies. In the first, the purpose is to identify market mechanisms allowing firms to relax price competition without being spatially separated. From this

perspective, the most natural approach is to assume that firms sell products that are differentiated in the space of characteristics. It combines both spatial and product differentiation per se. An alternative approach, however, is to appeal to some form of collusion between firms that permits them to avoid the devastating effects of price competition. This is especially relevant when products can hardly be differentiated. The second research strategy is based on Stigler (1961) and develops the idea that consumers are imperfectly informed about the places where the existing varieties are made available. In such a context, consumers must undertake some search before finding a good match.

Product Differentiation and Collusion

Several papers have shown that firms selling differentiated varieties choose to agglomerate at the market centre when products are sufficiently differentiated, transportation costs borne by the consumers are low enough, or both (de Palma et al. 1985). This can be understood as follows. When consumers have different tastes and when residential locations and tastes are not correlated (or, alternatively, when individuals exhibit a love of variety), each firm supplies what is the best match for consumers who are otherwise dispersed across all locations. Price competition is relaxed by product differentiation, so that firms may afford to set up at the place offering the best accessibility to their potential customers. Such a place is obviously the market centre when the consumer distribution along Main Street is uniform. In addition, it is never profitable for a firm to leave the cluster when transportation costs are low because the benefit of a good match dominates the additional transportation costs that the consumer must bear to buy her best match. All of this seems to fit modern economies characterized by more and more variety and decreasing travel costs. In a nutshell, we may then safely conclude that one of the main reasons for agglomeration is that *firms substitute product differentiation for spatial separation*, very much as *Newsweek* and *Time* are supplied in the same stores but differentiated by their cover stories (Irmen and Thisse 1998).

The welfare analysis of such an outcome is somewhat unexpected. At the optimum, prices are set equal to the common marginal cost so that consumers' wellbeing depends only upon firms' locations. In the case of a homogeneous good, maximizing total welfare boils down to minimizing aggregate transportation costs. However, once we introduce differentiation across varieties, consumers no longer patronize the nearest firm on each trip because they now benefit from intrinsic differentiation between stores. In this context, one needs a more general approach accounting for both distance and product diversity effects, the appropriate measure being the consumers' indirect utility. As a result, the formation of a cluster need not be socially sub-optimal. Quite the opposite: when products are sufficiently differentiated, transportation costs are low, or both, it is socially desirable to have all firms agglomerated within a cluster. Hence, unlike what Hotelling thought, such an extreme concentration may be socially optimal.

Under what became known as 'semi-collusion', it has been shown that firms that anticipate some form of collusion in the price stage, which is typically repeated, will choose to locate together at the market centre (Jehiel 1992; Friedman and Thisse 1993). In this case, selling a homogenous product makes it easier to sustain price collusion because the punishment for a defecting firm is more severe. Of course, collusion is not easy to maintain in the long run, so that firms face a positive probability that price collusion will break down. In this case, firms select separated locations but do not seek to maximize their spatial differentiation. Specifically, Jehiel et al. (1995) have established that the higher the probability that the price agreement will break down, the larger is the distance between firms.

Search

When firms sell differentiated products, it is reasonable to assume that consumers are incompletely informed about the varieties that are supplied. Even though the typical consumer knows which varieties are available in the market, she is unsure about which variety is offered where (and at which price). If consumers have to

compare alternatives before buying, they must undertake *search* among firms. Stated differently, when the only way for consumers to find out which variety is on offer in a particular store is to visit this store, they must bear the corresponding travel cost. Gathering information being costly, each consumer must compare the cost of an additional bit of information with the expected gain in terms of surplus. In a spatial setting, both the cost and the gain vary with consumers' and firms' locations.

When several stores are located together, it is reasonable to assume that the typical consumer knows the location and size of the cluster but not its composition. Once she arrives at the cluster, the travel costs are sunk and she can visit any store at a very low cost. But she must pay the transportation cost to each isolated store she visits. Spatial clustering of stores is, therefore, a particular means by which firms can facilitate consumer search. Indeed, a consumer is more likely to visit a cluster of stores than an isolated one because of the higher probability she faces of finding there a good match and a good price. When firms realize this fact, each of them understands that it might be in its own interest to form a marketplace with others. When a firm considers the possibility of joining competitors within the same marketplace, it thus faces a trade-off between a negative competition effect and a positive market area effect, both being generated by the pooling of firms selling similar products.

In the case of a market with a fixed size, Wolinsky (1983) has shown that the market outcome involves all firms forming a single marketplace once transportation costs are sufficiently low and when there are enough stores to make the cluster attractive. It is worth noting that the agglomeration may arise away from the market centre. Any point such that no single firm is able to find an alternative location far enough to induce some consumers to visit it before the cluster is a spatial equilibrium. Of course, the cluster cannot be too far from the market centre because stores need to offer a good accessibility to *all* consumers. Accordingly, once the urban area extends far away into the same direction, this implies that some firms will want to create a new cluster away from the original one.

Schulz and Stahl (1996) show that it is possible to uncover additional and surprising results by considering a market of variable size. To this end, they consider an unbounded space that allows them to capture the idea that more competition within the cluster may attract new customers coming from more distant locations, thus allowing the demand for each variety to increase. More concretely, the entry of a new variety may lead to an increase in the cluster's demand that outweighs the decrease in market area inflicted on existing varieties. Although price competition becomes fiercer, it appears here that firms may take advantage of the extensive margin effect to increase their prices in equilibrium. Clearly, when the number of varieties is not too large, such positive effects associated with the gathering of firms strengthen the agglomeration force that lies behind the cluster. Though collectively several firms might want to form a new market, it may not pay an individual firm to open a new market in the absence of a coordinating device. Consequently, a new firm entering the market will choose instead to join the incumbents, thus leading to a larger agglomeration. In this case, *the entry of a new firm creates a positive externality for the existing firms by making total demand larger*. This in turn explains the common fact that department stores encourage the location of competing firms within the shopping centre.

The Relationship with New Economic Geography

It appears that location theory and new economic geography have a lot in common, a fact that has been overlooked in the literature. Such a relationship between the two domains is worth noting because economic geography models are developed in general equilibrium frameworks involving monopolistic competition on the product market, whereas location theory uses partial equilibrium models under oligopolistic competition. Indeed, one of the main conditions identified in spatial competition for a cluster of firms to emerge

corresponds with the main finding established in 'new economic geography', that is, firms agglomerate when trade costs are sufficiently low (Krugman 1991; Fujita et al. 1999). Likewise, product differentiation fosters agglomeration whereas, by its mere existence, a cluster generates a lock-in effect similar to those encountered in economic geography. In both settings, the absence of increasing returns would lead to the emergence of 'backyard capitalism' in which each household produces its own consumption bundle. Finally, the market size effect uncovered in search models is similar to the agglomeration effect identified by Krugman and others.

Spatial Competition and Urban Economics

So far, consumers have been able to seek where to buy but not where to live. Yet it is reasonable to assume that consumers adjust their residential choices to the locations selected by large firms and/or by public facilities. For the resulting distribution to be non-degenerate, a land market must be introduced in which consumers compete for land use. In such a context, the demand of a consumer for the firm's output becomes in turn endogenous in that it depends on the income left after the land rent is paid. This brings into the picture some general equilibrium ingredients in that firms and households locations are interdependent. Fujita and Thisse (1986) consider a setting in which firms choose their locations, anticipating consumers' residential choices, this sequence reflecting the fact that firms have market power whereas consumers adjust their locational choices to those made by firms. Because they compete for land, consumers are spread around firms in a way such that no consumer can find a better place to live. In the case of two firms selling a homogeneous good at a common given price, the agglomeration of the two firms is always a Nash equilibrium. However, dispersed equilibria may also coexist when travel costs are sufficiently high. This is because the decrease in individual consumption resulting from a move toward the rival dominates the market area effect. In other words, firms may not find it advantageous to agglomerate, thus showing that *competition for land acts a major dispersion force*.

Public Facilities

Cities provide a large variety of local public goods. Because its location interacts with the locational choices of firms and households, a large public facility which consumers wish to access influences the nature of the urban structure. In particular, one expects the presence of a major equipment to act as an agglomeration force on the private sector (Thisse and Wildasin 1992). When topographical boundaries have no impact on the location of the public facility, this one is always established at the centre of the urban area and there is a tendency for this facility to draw the private firms together as income rises with respect to transportation costs. When the facility is set up near the edge of the area available for urban use – think of an urban area on the coast of a body of water – the resulting asymmetry has a significant impact on the locational interactions between firms: the two private firms are located together at the centre of the urban area. Hence, *the public facility may serve as the center of a dispersed spatial configuration, or it may induce the agglomeration of firms in a location different from that of the public facility itself*. In both cases, it vastly contributes to the shaping of the city structure.

Local Labour Markets

Due to the evolution of technological progress and the concomitant expansion of metropolitan areas, the urban labour force has become more heterogeneous whereas the labour market has been segmented in thinner sub-markets. The force inducing the formation of local labour markets finds its origin, at least partially, in the skill and geographical heterogeneity of workers (Brueckner et al. 2002). When workers have heterogeneous skills, firms have different job requirements because they have incentives to differentiate their job offers in order to gain market power in the labour market. This in turn implies that the labour market works as an *oligopsony* in which firms with different skill needs and different urban locations

compete for mobile and skill-heterogeneous workers. In terms of urban economics, each firm may be considered as a company town attracting workers who also choose to reside near this firm. As in the case of firms selling consumption goods, firms are separated in the geographical space because this allows them to enjoy market power over the workers situated in their vicinity. Consequently, the economy may be viewed as a system of cities in which each firm/city competes to attract workers who are also residents. *The fact that each firm is anchored in a distinct location is a fundamental reason for the emergence of local labour markets*.

When workers bear the training cost that allows them to erase any mismatch between their innate skills and the skill needs of their employer, the net wage is lower for workers whose 'skill distance' from their employer is larger. Firms understand that, in the residential equilibrium, commuting distance is positively related to a worker's skill distance from the firm. In such a context, the equilibrium residential location of workers is governed by the quality of their match in the labour market. Knowledge of the connection between skill and commute distances affects the firm's interaction with its rivals as it competes for labour. The critical issue is that the equilibrium wage depends on the commuting cost parameters, yielding a link between the urban structure and the labour market. More precisely, low-skill workers incur high commuting costs, which may in turn lead low-skill workers not to take a job. Unemployment may arise, therefore, because some workers turn out to be too 'distant' from firms in *both* the skill and urban spaces. As in the foregoing, two different spatial components interact to shape the social structure of cities.

See Also

- ▶ New Economic Geography
- ▶ Product Differentiation
- ▶ Spatial Economics
- ▶ Systems of Cities
- ▶ Urban Agglomeration
- ▶ Urban Economics

Bibliography

Aumann, R. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Brueckner, J., J.-F. Thisse, and Y. Zenou. 2002. Local labor markets, job matching and urban location. *International Economic Review* 43: 155–171.

d'Aspremont, C., J. Gabszewicz, and J.-F. Thisse. 1979. On Hotelling's 'stability in competition'. *Econometrica* 47: 1045–1050.

de Palma, A., V. Ginsburgh, Y.Y. Papageorgiou, and J.-F. Thisse. 1985. The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica* 53: 767–781.

Friedman, J., and J.-F. Thisse. 1993. Partial collusion fosters minimum product differentiation. *RAND Journal of Economics* 24: 631–645.

Fujita, M., P. Krugman, and A. Venables. 1999. *The spatial economy: Cities, regions and international trade*. Cambridge, MA: MIT Press.

Fujita, M., and J.-F. Thisse. 1986. Spatial competition with a land market: Hotelling and von Thünen unified. *Review of Economic Studies* 53: 819–841.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Hurter, A., and J. Martinich. 1988. *Facility location and the theory of production*. Dordrecht: Kluwer.

Irmen, A., and J.-F. Thisse. 1998. Competition in multi-characteristics spaces: Hotelling was almost right. *Journal of Economic Theory* 78: 76–102.

Jehiel, P. 1992. Product differentiation and price collusion. *International Journal of Industrial Organization* 10: 633–641.

Jehiel, P., J. Friedman, and J.-F. Thisse. 1995. Collusion and antitrust detection. *The Japanese Economic Review* 46: 226–246.

Koopmans, T. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99: 483–499.

Rosen, S. 2002. Markets and diversity. *American Economic Review* 92: 1–15.

Schulz, N., and K. Stahl. 1996. Do consumers search for the highest price? Equilibrium and monopolistic optimum in differentiated products markets. *RAND Journal of Economics* 27: 542–562.

Selten, R. 1965. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft* 121: 301–324.

Stigler, G. 1961. The economics of information. *Journal of Political Economy* 69: 213–225.

Thisse, J.-F., and D. Wildasin. 1992. Public facility location and urban spatial structure. *Journal of Public Economics* 48: 83–118.

Tirole, J. 1988. *The theory of industrial organization*. Cambridge, MA: MIT Press.

Weber, A. 1909. *Ueber den standort der industrien*. Tübingen: J.C.B. Mohr. English translation: *The*

Theory of the Location of Industries. Chicago: Chicago University Press, 1929.
Wolinsky, A. 1983. Retail trade concentration due to consumers' imperfect information. *Bell Journal of Economics* 14: 275–282.

Locke, John (1632–1704)

Karen I. Vaughn

John Locke, the philosopher and author of *Essay Concerning Human Understanding, Two Treatises of Government*, and *A Letter Concerning Toleration*, was educated at Westminster School and Christ Church, Oxford, where he received a BA in 1656 and an MA in 1659. He lectured in Greek and Moral Philosophy, studied experimental medicine on his own initiative, and attended Robert Boyle's unorthodox experimental group in his spare time. In 1666 he joined the household of Anthony Ashley Cooper, the first Earl of Shaftesbury, where he developed an interest in political and economic matters. Locke's cautious political involvements caused him to spend a brief period of exile in Holland during the 1680s until James II abdicated and William and Mary ascended the British throne. Locke was known as both a philosopher and a public servant. Among his contributions to public life was his organizing of the Board of Trade in 1695 and his subsequent service as a commissioner of the Board until 1700.

Locke's contributions to economic thought consisted of two major essays and a minor pamphlet. Locke's first economic essay, *Some Considerations of the Consequences of the Lowering of Interest, and Raising the Value of Money*, was also his most important. Although he worked on drafts of this essay for more than twenty years, he only published it in 1691 in an attempt to influence Parliament to defeat a bill to lower the legal rate of interest from 6 per cent to 4 per cent. Although the essay, being too complicated and analytic for successful polemic, failed to persuade Parliament, its very failures as political persuasion make it an interesting essay in the history of economic thought.

In an age of mercantilist confidence in the importance of economic regulation, Locke, the natural law philosopher, began his essay with the question, 'Whether the price of the hire of money (the rate of interest) can be regulated by law?' (*Some Considerations*, p. 1). His answer was that 't' is manifest it cannot' because interest is a price and all prices are determined by laws of nature that are beyond the control of mere political law. Politicians can pass interest rate legislation, but they are powerless to effect the results they intend. People unwilling to give up the 'chance of gain' will evade the interest rate ceilings in ways that will drive the effective interest rate higher than the market rate would have been in the absence of legislation and cause shortages of loanable funds, hamper trade and redistribute wealth in undeserved ways.

Locke based his conclusions on a carefully developed theory of price determination which he applied to all exchangeable goods. His price theory, while primitive by modern standards, was nevertheless remarkably accurate in predicting the correct direction of change of price in response to changes in underlying variables. He clearly showed that quantity demanded is inversely related to price, while quantity supplied is directly related. He had the concept of an equilibrium price and in fact used the term equilibrium in several contexts. The fact that the price with which he was most concerned was the interest rate also led him to develop a very advanced monetary theory, probably his most important contribution to the development of economic thought.

Locke's theory of value starts from the premise that only relative values are important to economic exchange. There are no intrinsic values that make some quantity of one good always exchange for some quantity of another good. Exchange value depends exclusively upon the proportion between the quantity of a good offered for sale and its vent, the quantity of the good demanded. The same principles applied to the value of money. Money was a special good in that it had two values: a value in use and a value in exchange. Its value in use was as money capital and its price was the rate of interest.

The rate of interest, then, depended not upon legislation but upon the profitability of investment and the alternative uses of money to individuals. The value of money in exchange, however, was a special case of the value of goods because while its quantity was variable, its vent was 'always sufficient' (*Some Considerations*, p. 71), by which Locke meant that the public would be willing to hold any amount of money in circulation. Hence, he argued, the value of money, or its purchasing power was solely an inverse function of the quantity of money in circulation. The quantity of money, however, also depended upon its 'quickness of circulation' (ibid., p. 33), which he analysed as a function of the relatively stable payments habits of the community. It is easy to see an early statement of the quantity theory in Locke's analysis.

Locke applied his quantity theory of money to the problem of international price levels and specie flows. His analysis here is interesting because he argued that 'any quantity of money . . . would serve to drive any proportion of trade' (ibid., pp. 75–6 and hence the absolute amount of specie in any one country is irrelevant to its welfare. However, he also believed that countries which were involved in world trade needed to maintain a certain proportion between the quantity of money and the volume of trade so that it would not be disadvantaged relative to its trading partners. While this position is inconsistent with a simple quantity theory of money (as Hume was to point out in the next century), Locke assumed that unfavourable terms of trade would lead to depression and the emigration of skilled labour. In other words, he assumed output effects consequent on changes in the money supply that made him reject an untrammelled reliance on the price-specie-flow mechanism to maintain a balance of trade.

Locke's second major essay, *Further Considerations Concerning Raising the Value of Money* (1695), while it reiterated many of his earlier arguments, was mostly concerned with the issue of recoinage. Locke took issue with a proposal to devalue the official coinage by 20 per cent and argued strongly for recoining at the old standard the currently circulating coins debased by clipping and normal wear and tear. Money, Locke argued was equivalent to gold and silver. People contracted for gold and silver and a government stamp was simply an assurance of the specie content of official coins. Hence, a devaluation would only confuse trade and cause an increase in prices denominated in terms of pounds and shillings.

It has been pointed out that a 20 per cent official devaluation would simply have ratified a de facto devaluation that had taken place through clipping and wear and tear, and so Locke's arguments were not pertinent to the problem. While that may well have been correct, Locke's real agenda was to argue that money was the private property of citizens and not a creation of government. He was making a moral argument that was consistent with the political philosophy he espoused in his *Second Treatise of Government*.

From the point of view of the economist, Locke's *Second Treatise* is interesting primarily because of the theory of property developed therein. Locke argues that individuals earn the right to private property by virtue of the fact that in order to survive, they must mix their own labour with unowned resources. Locke goes on to argue that one has a right to create as much property (in land as well as in goods) as one can for one's use as long as nothing is wasted and as long as there are sufficient resources left for other also to make a living. From this basis, Locke developed a theory of the origin of money where money arises out of man's efforts to provide a store of value to prevent waste and a theory of the origin of the state based on scarcity of land.

In the course of making his point, Locke argued that private property is not only moral, it is also practical since labour is productive and accounts for 99/100 of the value of things 'useful to the life of man' (*Second Treatise*, p. 314). While this statement was only illustrative and was not meant to suggest any causal relationship between labour input and market price, it did provide a starting point for later labour theories of value.

Selected Works

1690. *Two treatises of government*, ed. Peter Laslett, 2nd ed. Cambridge: Cambridge University Press, 1953.

1696. Several papers relating to money, interest and trade, etcetera. New York: Augustus M. Kelley, 1968.

Bibliography

Laslett, P. 1957. John Locke, the great recoinage, and the origins of the board of trade, 1695–1698. *William and Mary Quarterly* 14: 370–392.
Leigh, A.H. 1974. John Locke and the quantity theory of money. *History of Political Economy* 6: 200–219.
Letwin, W. 1964. *The origins of scientific economics*. Garden City: Anchor Books.
Vaughn, K.I. 1980. *John Locke: Economist and social scientist*. Chicago: University of Chicago Press.

Logit Models of Individual Choice

Thierry Magnac

Abstract

The logit model was named by Berkson after probit, its close competitor; the two are the most popular econometric methods used in applied work to estimate models for binary variables. It can be easily extended to the treatment of multinomial variables and enjoys specific properties in panel data binary models. Increasingly flexible logit models have also been elaborated for demand analyses. Their development has been stimulated by the increasing availability of databases on individual discrete choices. Because generalized logit models belong to the class of random utility models, their use has promoted sound applied economic research in demand analysis.

Keywords

Asymptotic least squares; Conditional likelihood; Discrete choices; General extreme value distributions; Generalized linear models; Incidental parameters; Independence of irrelevant alternatives; Logit models of individual choice; Maximum likelihood; Method of moments; Minimum distance; Mixture models; Probit models; Random utility models; Simulation methods; Spatial statistics; Statistical mechanics; Two-level nested logit

JEL Classifications

C25; C35; D11

The logit function is the reciprocal function to the sigmoid *logistic* function. It maps the interval [0,1] into the real line and is written as:

$$logit(p) = \ln(p/(1-p)).$$

Two traditions are involved in the modern theory of logit models of individual choices. The first one concerns *curve fitting* as exposed by Berkson (1944), who coined the term 'logit' after its close competitor 'probit' which is derived from the normal distribution. Both models are by far the most popular econometric methods used in applied work to estimate models for binary variables, even though the development of semiparametric and nonparametric alternatives since the mid-1970s has been intensive (Horowitz and Savin 2001).

In the second strand of literature, models of discrete variables and discrete choices as originally set up by Thurstone (1927) in psychometrics have been known as 'random utility models' (RUM) since Marschak (1960) introduced them to economists. As the availability of individual databases and the need for tools to forecast aggregate demands derived from discrete choices were increasing from the 1960s onwards, different waves of innovations, fostered by McFadden (see his Nobel lecture, 2001) elaborated more and more sophisticated and flexible logit models. The use of these models and of simulation methods has triggered burgeoning applied research in demand analysis in recent years.

Those who wish to study the subject in greater detail are referred to Gouriéroux (2000), McFadden (2001) or Train (2003), where references to applications in economics and marketing can also be found.

Measurement Models

As Berkson (1951, p. 327) put it, logit (or probit) models may be seen as 'merely a convenient way of graphically representing and fitting a function'. They are used for any empirical phenomenon delivering a binary random variable Y_i, taking values 0 and 1, to be analysed. In a logit model, it is postulated that its probability distribution conditional on a vector of covariates X_i is given by:

$$\Pr(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

where β is a vector of parameters. This model can also be derived from more general frameworks in statistical mechanics or spatial statistics (Strauss 1992).

With the use of cross-sectional samples, the parameter of interest is estimated using maximum likelihood or by generalized linear models (GLM) methods where the link function is logit (McCullagh and Nelder 1989). Under the maintained assumption that it is the true model and other standard assumptions, the maximum likelihood estimator (MLE) is consistent, asymptotically normal and efficient (Amemiya 1985). Nevertheless, the MLE may fail to exist, or more exactly be at the bounds of the parameter space, when the samples are uniformly composed of 0 s or 1 s, for instance (Berkson 1955).

When repeated observations are available, the method of Berkson delivers an estimator close to MLE since they are asymptotically equivalent. Observe first that the logit function of the true probability obeys the linear equation:

$$logit(\Pr(Y_c = 1|X_c)) = X_c\beta$$

where the covariates X_c now take a discrete number of values defining each cell, c. Second, use the observed frequency in each cell, $\hat{p}_c$, and contrast it with the theoretical probability, p_c, as:

$$\begin{aligned} logit(\hat{p}_c) &= X_c\beta + (logit(\hat{p}_c) - logit(p_c)) \\ &= X_c\beta + \varepsilon_c. \end{aligned}$$

The random term ε_c properly scaled by the square root of the number of observations in cell c is asymptotically normally distributed with variance equal to $1/(p_c(1 - p_c))$. The method of Berkson then consists in using minimum chi-square, that is, a method of moments, to estimate β, an instance of what is know as minimum distance or asymptotic least squares (Gouriéroux et al. 1985).

When measurements for a single individual are repeated, Rasch (1960) suspected that individual effects might be important and proposed to write:

$$logit(\Pr(Y_{it} = 1|X_{it})) = X_{it}\beta + \delta_i$$

where t indexes the different items that are measured and δ_i is an individual specific intercept or fixed effect. Items can be different questions in performance tests or different periods. In the original Rasch formulation, parameters were allowed to be different across items, β_t, and there were no covariates.

Given that the number of items is small, it is well known that the estimation of such a model runs into the problem of incidental parameters (see Lancaster 2000). As the number of parameters δ_i increases with the cross-section dimension, the MLE is inconsistent (Chamberlain 1984). Nevertheless, the nuisance parameters δ_i can be differenced out using conditional likelihood methods (Andersen 1973) because:

$$\begin{aligned} &logit(\Pr(Y_{it} = 1|X_{it}, Y_{it} + Y_{it'} = 1)) \\ &\quad = (X_{it} - X_{it'})\beta. \end{aligned}$$

The conditional likelihood estimator of β is consistent and root n asymptotically normal but it is not efficient, although no efficient estimator is known. Furthermore, when binary variables Y_{it} are independent, conditionally on X_i, the only model where a root n consistent estimator exists is a logit model (Chamberlain 1992). Extensions of Rasch rely on the fact that root n consistent estimators exist if and only if $Y_{it} + Y_{it'}$ is a sufficient statistic for the nuisance parameters δ_i (Magnac 2004). When the number of items or periods becomes large, profile likelihood methods where individual

effects are treated as parameters seem to be accurate in Monte Carlo experiments as soon as the number of periods is four or five (Arellano 2003).

Multinomial logit (or in disuse 'conditional logit') is to binary logit what a multinomial is to a binomial distribution (Theil 1969). Given a vector Y_i consisting of K elements which are binary random variables and lie in the $\mathbb{R}^K$ – simplex (their sum is equal to 1), it is postulated that:

$$\Pr\left(Y_i^{(k)} = 1 | X_i\right) = \frac{\exp\left(X_i\beta^{(k)}\right)}{1 + \sum_{k=2}^{K} \exp\left(X_i\beta^{(k)}\right)}$$

where by normalization, $\beta^{(1)} = 0$. Ordered logit has a different flavour since it applies to rank-ordered data such as education levels (Gouriéroux 2000).

As probits, logit models are very tightly specified parametric models and can be substantially generalized. Much effort has been exerted to relax parametric and conditional independence assumptions, starting with Manski (1975). Manski (1988) analyses the identifying restrictions in binary models, and Horowitz (1998) reviews estimation methods. In some cases, Lewbel (2000) and Matzkin (1992) offer alternatives.

Random Utility Models

The theory of discrete choice is directly set up in a multiple alternative framework. A choice of an alternative k belonging to a set C is assumed to be probabilistic either because preferences are stochastic or heterogenous, or because choices are perturbed in a random way. By definition, choice probability functions map each alternative and choice sets into the simplex of $\mathbb{R}^K$.

A strong restriction on choices is the axiom of Independence of Irrelevant Alternatives (IIA, Luce 1959). The axiom states that the choice between two alternatives is independent of any other alternative in the choice set. The version that allows for zero probabilities (McFadden 2001) states that for any pair of choice set C, C' such that $\{k, k'\} \in C$ and $C \subset C'$:

$\Pr(k \text{ is chosen in } C') = \Pr(k \text{ is chosen in } C).\Pr(\text{An element of } C \text{ is chosen in } C')$.

Under this axiom, choice probabilities take a multinomial generalized logit form.

Moreover, assume that choices are associated with utility functions, $\{u^{(k)}\}_k$ that depend on determinants X_i and random shocks:

$$u^{(k)} = X\beta^{(k)} + \varepsilon^{(k)},$$

and that the actual choice of the decision maker yields maximum utility to her. Then, the IIA axiom is verified if and only if $\varepsilon^{(k)}$ are independent and extreme value distributed (McFadden 1974). Extensions of decision theory under IIA were proposed in the continuous case (Resnick and Roy 1991) or in an intertemporal context (Dagsvik 2002).

The IIA axiom is a strong restriction as in the famous red and blue bus example where, if IIA is assumed, the existence of different colours affects choices of transport between bus and other modes while introspection suggests that colours should indeed be irrelevant. Several generalizations which proceed from logit were proposed to bypass IIA. Hierarchical or tree structures were the first to be used. At the upper level, the choice set consists of broad groups of alternatives. In each of these groups, there are various alternatives which can consist themselves of subsets of alternatives, and so on. The best-known model is the two-level nested logit, where alternatives are grouped by similarities. For instance, the first level is the choice of the type of the car, the second level is the make of the car. The formula of choice probabilities for nested logit,

$$p^{(k)} = \frac{\exp\left(X\beta^{(k)}/\lambda_{B_s}\right)\left(\sum_{j\in B_s}\exp\left(X\beta^{(j)}/\lambda_{B_s}\right)\right)^{\lambda_{B_s}-1}}{\sum_{t=1}^{T}\left(\sum_{j\in B_t}\exp\left(X\beta^{(j)}/\lambda_{B_s}\right)\right)^{\lambda_{B_t}}},$$

where alternative k belongs to B_s, is not illuminating but the logic of construction is clear. Choices at each level are modelled as multinomial logit (Train 2003).

General extreme value distributions (McFadden 1984) provide more extensions, although they do not generate all configurations of choice probabilities. In contrast, mixed logit does, as shown by McFadden and Train (2000). Instead of considering that parameters are deterministic, make them random or heterogeneous across agents. The result is a mixture model where individual probabilities of choice are obtained by integrating out the random elements as in

$$p^{(k)} = \int p^{(k)}(\beta) f(\beta) d\beta.$$

Integrals are computed using simulation methods (MacFadden 2001). The same principle is used by Berry et al. (1995) with a view to generalizing the aggregate logit choice models using market data. Logit models are still very much in use in applied settings in demand analysis and marketing, and are equivalent to a representative consumer model (Anderson et al. 1992). Mixed logits permit much more general patterns of substitution between alternatives and should probably become the standard tool in the near future.

See Also

- Categorical Data
- Econometrics
- Hierarchical Bayes Models
- Maximum Likelihood
- McFadden, Daniel (born 1937)
- Mixture Models
- Non-linear Panel Data Models
- Product Differentiation
- Rational Behaviour
- Utility

Bibliography

Amemiya, T. 1985. *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Andersen, E.B. 1973. *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forlag.

Anderson, S.P., A. de Palma, and J.F. Thisse. 1992. *Discrete choice theory of product differentiation*. Cambridge, MA: MIT Press.

Arellano, M. 2003. Discrete choices with panel data. *Investigaciones Economicas* 27: 423–458.

Berkson, J. 1944. Application of the logistic function to bioassay. *Journal of the American Statistical Association* 39: 357–365.

Berkson, J. 1951. Why I prefer logits to probits. *Biometrics* 7: 327–339.

Berkson, J. 1955. Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association* 50: 130–162.

Berry, S.T., J.A. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890.

Chamberlain, G. 1984. Panel data. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligator, Vol. 2. Amsterdam: North-Holland.

Chamberlain, G. 1992. *Binary response models for panel data: Identification and information*. Cambridge: Harvard University.

Dagsvik, J. 2002. Discrete choice in continuous time: Implications of an intertemporal version of IAA. *Econometrica* 70: 817–831.

Gouriéroux, C. 2000. *Econometrics of qualitative dependent variables*. Cambridge: Cambridge University Press.

Gouriéroux, C., A. Monfort, and A. Trognon. 1985. Moindres carrés asymptotiques. *Annales de l'INSEE* 58: 91–121.

Horowitz, J. 1998. *Semiparametric methods in econometrics*. Berlin: Springer.

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

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95: 391–413.

Lewbel, A. 2000. Semiparametric qualitative response model estimation with unknown heteroskedasticity or instrumental variables. *Journal of Econometrics* 97: 145–177.

Luce, R. 1959. *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Magnac, T. 2004. Panel binary variables and sufficiency: Generalizing conditional logit. *Econometrica* 72: 1859–1877.

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

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

Marschak, J. 1960. Binary choice constraints and random utility indicators. In *Mathematical methods in the social sciences*, ed. K. Arrow. Stanford: Stanford University Press.

Matzkin, R. 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60: 239–270.

McCullagh, P., and J.A. Nelder. 1989. *Generalized linear models*. London: Chapman and Hall.
McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic Press.
McFadden, D. 1984. Econometric analysis of qualitative response models. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, Vol. 2. Amsterdam: North-Holland.
McFadden, D. 2001. Economic choices. *American Economic Review* 91: 351–378.
McFadden, D., and K. Train. 2000. Mixed MNL models for discrete responses. *Journal of Applied Econometrics* 15: 447–470.
Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark Paedagogiske Institut.
Resnick, S.I., and R. Roy. 1991. Random USC functions, max stable process and continuous choice. *Annals of Applied Probability* 1: 267–292.
Strauss, D. 1992. The many faces of logistic regression. *American Statistician* 46: 321–327.
Theil, H. 1969. A multinomial extension of the linear logit model. *International Economic Review* 10: 251–259.
Thurstone, L. 1927. A law of comparative judgement. *Psychological Review* 34: 273–286.
Train, K. 2003. *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.

Logit, Probit and Tobit

Forrest D. Nelson

Two convenient classifications for variables which are not amenable to treatment by the principal tool of econometrics, regression analysis, are *quantal responses* and *limited responses*. In the quantal response (all or nothing) category are dichotomous, qualitative and categorical outcomes, and the methods of analysis identified as *probit* and *logit* are appropriate for these variables. Illustrative applications include decisions to own or rent, choice of travel mode, and choice of professions. The limited response category covers variables which take on mixtures of discrete and continuous outcomes, and the prototypical model and analysis technique is identified as *tobit*. Examples are samples with both zero and positive expenditures on durable goods, and models of markets with price ceilings including data with both limit and non-limit prices. While the tobit model evolved out of the probit model and the limited and quantal response methods share many properties and characteristics, they are sufficiently different to make separate treatment more convenient.

Dichotomous Logit and Probit Models

The simplest of the logit and probit models apply to dependent variables with dichotomous outcomes. If Y can take on only two possible outcomes, say 0 and 1, then the stochastic behaviour of Y is described by the probability of a positive response, $P(Y = 1|X)$, which is here taken to depend on a vector valued variable X. The specification of the functional form for P in the probit model is the normal CDF, while the logit model uses the logistic equation. Specifically, for the probit model

$$P(Y = 1|X = \int_{-\infty}^{\theta' X} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-1}{2}u^2\right] \mathrm{d}u = \Phi(\theta' X) \tag{1}$$

while for the logit model

$$P(Y = 1|X = \frac{\exp(\theta' X)}{1 + \exp(\theta' X)} \equiv \Lambda(\theta' X). \tag{2}$$

In both forms, θ is a parameter vector, and the choice of a linear, additive form for the way X enters Φ and Λ is common but not necessary.

A rudimentary derivation of these otherwise ad hoc specifications from an explicit description of behaviour is as follows. Suppose the underlying theory of behaviour posits a continuous but latent (or observable) variable, say W, embodying that behaviour, with the dichotomous realization on Y determined by comparing W with some 'threshold'. For convenience, take the threshold to be zero. Then Y is determined by

$$Y = \begin{cases} 1 \text{ if } W > 0 \\ 0 \text{ if } W \leq 0 \end{cases}$$

If W is related to X linearly with an additive random component,

$$W = \beta' X - u,$$

then outcome probabilities are determined by

$$P(Y = 1|X) = P(\beta' X - u > 0) = F\left(\frac{\beta' X}{\delta}\right), \quad (3)$$

where δ is a scale parameter used to standardize the random variable u, and F is the CDF for this standardized random variable u/δ. The choice of F is dictated by the distribution of u. If u/δ is $N(0,1)$, then (3) is the probit model of equation (1); and if u/δ is logistic with mean 0 and variance $\pi^2/3$, then (3) is the logit model of equation (2). In both cases, $\theta = \beta/\delta$, and β and δ are not separately identifiable.

The similarities in the shapes of the logistic and normal distributions suggest that results of probit and logit analysis will differ by very little. Indeed, the inferences drawn from the two methods applied to the same data are invariably similar, and even parameter estimates from the two models will agree, approximately, up to a factor of proportionality. (Logit coefficients tend to exceed probit coefficients by a scale factor in the range 1.6 to 1.8.) A choice between the two models, therefore, is not an important one and may often be ruled by convenience factors, such as availability of appropriate computer programs.

One superficially attractive alternative to probit or logit is the linear probability model (LPM). The LPM specification is

$$P(Y = 1, |X) = \theta' X. \quad (4)$$

Since in the dichotomous case $E(Y|X) = P(Y = 1|X)$, this model lends itself to a simple linear regression formulation,

$$Y = \theta' X + u \quad (5)$$

If (4) is correct, then the disturbance term u defined implicitly in (5) has mean zero and is uncorrelated with X. Least squares would thus appear to be a viable estimator for θ. But least squares does not recognize the implicit constraint in (4) that $\theta' X$ must lie in the interval from zero to one. Indeed this constraint makes the linear form unattractive except as a local approximation, and, if (4) is merely an approximation, the independence of u and X will not hold, so least squares will produce biased estimates of even the linear approximation. Such problems generally outweigh the advantages of the linear specification, and sigmoid shapes for P, particularly the logit and probit models, are more commonly selected.

The most frequently used estimation technique for the dichotomous logit and probit models of equations (1) and (2) is maximum likelihood (ML). For a random sample of observations on the 0 –1 dependent variable Y_i and independent variables X_i, $i = 1, \ldots, N$, the maximum likelihood estimator, θ, is found as the solution to

$$\max_{\theta} \left\{ \ln L(\mathbf{Y}|\mathbf{X}, \theta) = \sum_{i=1}^{N} [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)] \right\}. \quad (6)$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)'$, $\mathbf{X} = (X_1, \ldots X_N)'$ and $P_i = \Phi(\theta' X_i)$ for the probit model or $P_i = \Lambda(\theta' X_i)$ for the logit model. The first order conditions $(\partial(\ln L)/\partial\theta = 0)$ are nonlinear in θ, so explicit solutions do not exist, and iterative maximization methods must be employed. Fortunately, the log likelihood is globally concave for probit and logit models, so any of a number of common algorithms will suffice. The Newton–Raphson algorithm, for example, is quite satisfactory and yields a consistent estimate of the covariance matrix of θ, the negative inverse of the matrix of second derivatives of the log likelihood, as a by-product.

An alternative estimation method is applicable when the data consist of replicated observations on Y. Suppose that corresponding to each of the N observations on X_i, there are n_i observations on Y, say $Y_{ij}, j = 1, \ldots, n_i$. A sufficient statistic for the ni observations Y_{ij} is given by the fraction of positive responses, $p_i = (\Sigma_j Y_{ij})/n_i$. Using the logit model as an example, define the "observed

logit" as $w_i = \Lambda^{-1}(p_i)$, note that the "true logit" is $\Lambda^{-1}(P_i) = \theta' X$, and let the difference between them be $u_i = w_i - \theta' X_i$. A Taylor series expansion of $\Lambda^{-1}(p_i)$ about P_i reveals that, for large enough ni, ui is approximately $N\{0,1/[n_i p_i(1-p_i)]\}$. Thus, weighted least squares estimation of θ from the regression equation,

$$w_i = \theta' X_i + u_i \tag{7}$$

yields an estimator $\widehat{\theta}$ with the same asymptotic properties as the MLE. In the logit model the transformation Λ^{-1} is the log-odds ratio, so $w_i = \ln[p_i/(1 - p_i)]$. The analysis is similar for the probit model. The estimator was first derived by Berkson (1944) using the estimation principle of minimum chi-square, and Theil (1969) obtained it for the more general multinomial case using the weighted least squares principle as described here. Thus the estimator is interchangeably referred to as the Berkson–Theil WLS estimator and the MIN chi-square estimator.

The ML and WLS estimators are both consistent, asymptotically efficient and asymptotically normal with the same covariance matrix, so there is little basis for choice between the two except for the computational advantages of WLS. Of course, WLS applies only to replicated data, and the two estimators differ in terms of sample size requirements – the properties of the MLE rely on the total number of observations, $\Sigma_i n_i$, while the asymptotic approximations of the WLS estimator are valid only if the number of replications, n_i, is large for each i. Both estimators have been criticized for lack of robustness against misspecification of the functional form of $P(Y = 1|X)$. Estimators which are robust against this misspecification have been proposed by Manski (1975) and Cosslett (1983). The basic idea is to restrict $P(Y = 1|X) = F(\theta' X)$ only so far as to require F to be monotonic and estimate the parameters θ and the function F simultaneously.

These simple models have been extended in a number of ways, and they are closely related to a number of other analysis techniques. (Amemiya (1981) provides a comprehensive survey). The most obvious extension is to allow Y to take on more than two values – the resulting extension for the logit model is referred to as multinomial logit, while the corresponding extension for probit turns out to be computationally onerous for more than four alternatives. McFadden obtained a multinomial logistic form for probabilistic discrete choice behaviour from a random utility model which incorporates such additional features as alternative specific attributes (McFadden 1974). Models with multiple, ordered outcomes for the dependent variable have also been proposed (McKelvey and Zavoina 1975).

While probit and logit models specify the conditional distribution of Y given X, discriminant analysis begins with a specification of the conditional distribution of X given Y. Interestingly, the implied form for $P(Y \mid X)$ in the normal discriminant analysis model with two populations is the same as equation (2), but the differences in assumptions made under the two approaches lead to different estimators with different properties. In a similar comparison, the log-linear models for contingency tables (see, for example, Bishop et al. 1975) specify a functional form for the joint distribution of a set of qualitative variables. It is easy to show that the implied conditional probability of one of these variables, conditional on the rest, has the multinomial logistic form.

Tobit Model

In the standard regression model, the dependent variable is generally assumed to take on any of an infinite continuum of values, and the probability of any particular value is zero. In the dichotomous probit model, the dependent variable assumes only two values, each of which is assigned probability mass. Tobin (1958) proposed a limited variable model, later called the Tobit model, to handle dependent variables which are mixtures of these two cases, specifically mass points at the low end and continuous values above. An example is the analysis of durable goods expenditures when negative values cannot occur but there are frequent observations of zero. The specification of the model is

$$Y = \begin{cases} \beta' X + u & \text{if RHS} > 0, \\ 0 & \text{if RHS} \leq 0. \end{cases} \tag{8}$$

In the statistics literature, the term 'truncated' is applied to a univariate model in which there is no record of observations beyond the limit point, and the term 'censored' is used for situations in which the number of limit observations is recorded even though their values are not. Though the Tobit model is a multivariate one, it is closest to the censored variable case with the additional requirement that all exogenous variables be recorded for both limit and non-limit observations.

Even under the standard assumptions of the regression model, namely that the error terms u are independent of X and of each other, have zero mean, and are homoscedastic, it is easily seen that ordinary least squares regression of Y on X will lead to biased estimates of β. Observations with large negative disturbances u are more likely to be censored than are observations with large positive value of u, so the regression disturbance, defined as $\upsilon = Y - \beta' X$ for both limit and non-limit observations, will have a mean which is greater than zero and in fact depends on X. Thus, the least squares estimator of β will be biased, usually toward zero, and that is true whether or not limit observations are included in the sample.

An estimator with desirable asymptotic properties is maximum likelihood. Under the assumption that the error terms u are iid normal, the likelihood function is given by

$$L(\mathbf{Y}|\mathbf{X}, \beta, \sigma) = \prod^{\psi_0} \Phi\left(\frac{-\beta X_i}{\sigma}\right) \cdot \prod^{\psi_0} \times \frac{1}{\sigma}\phi\left(\frac{Y_i - \beta' X_i}{\sigma}\right), \tag{9}$$

where Ψ_0 and Ψ_1 are the sets of observation subscripts i corresponding to limit and non-limit observations, respectively, and Φ and φ are the standard normal distribution and density functions. The first order conditions for maximization of the logarithm of (9) are non-linear, so iterative techniques are required. Olsen (1978) pointed out that, under the reparameterization $(\alpha', \alpha_0) = (\beta'/\sigma, -1/\sigma)$ the log-likelihood is globally concave, so an algorithm such as Newton- Raphson will yield the MLE starting from any initial estimate of the parameters. Amemiya (1973) demonstrated the conditions under which the MLE will be consistent, asymptotically efficient and asymptotically normal.

It is widely recognized that the MLE for this Tobit model is not robust against misspecifications which would be innocuous in the corresponding regression model. (See Hurd (1979), Arabmazar and Schmidt (1981) and Goldberger (1983) for examples, Robinson (1982) for an exception involving serial correlation, and Nelson (1981) for a specification test.) This lack of robustness has stimulated the development of alternative estimators such as Powell's (1984) least absolute deviation estimator. These alternatives, unfortunately, are computationally difficult, do not carry over easily to related models, and are not in widespread use.

The Tobit model has been extended in various ways. Trivial adaptations include non-zero thresholds which are constant or at least exogenous and censoring from above rather than below. The truncated variable case with no limit observations is easily handled with modification of the likelihood function (Hausman and Wise 1977). Similar modifications allow for interior censoring and for both upper and lower truncation (Rosett 1959; Rosett and Nelson 1975).

Richer generalizations of the Tobit model involve multiple equations. Three examples are simultaneous equations models, models of markets in disequilibrium and models with self-selection. The first of these is a generalization of the simultaneous equation techniques for linear models. (Lee (1981) surveys and extends this literature.) In the second example, quantities supplied and demanded serve as the upper truncation points for each other, and the two equations are estimated simultaneously (see Quandt (1982) for a survey). In models with self-selection, one behavioural relation determines whether the dependent variable of a second equation will be observed. Heckman (1974) develops such a model in which the decision to participate in the

labour force and the level of participation are the two equations of interest.

An extensive treatment of limited dependent variable models can be found in Maddala (1983), and Amemiya (1984) provides a comprehensive yet compact survey.

Origin and Evolution of Probit, Logit and Tobit Analysis

Credit for invention of the method of analysis later to be called probit is generally given to the psychophysicist Fechner (1860). In assessing the ability of subjects to perceive differences in the weight of inanimate objects, Fechner converted the proportion of correct responses to normal deviates and plotted those deviates against the true weight difference. Urban (1909) collected extensive data on problems of this nature and introduced such extensions as the use of the Cauchy cdf in place of the normal curve.

Introduction of these quantal response techniques to bioassay came as early as the 1920s, the most influential early contributors being Gaddum and Bliss. (Biological assay is the measurement of the potency of some stimulus by means of the reactions which it produces in living matter.) Gaddum (1933) transformed the proportion of positive quantal responses into its 'normal equivalent deviation' (n.e.d.) by the inverse normal cdf. Then he fitted straight lines to the plot of this n.e.d. against the stimulus level. In parallel but independent research, Bliss (1934) applied the term 'probit', a contraction of 'probability unit', to the n.e.d. increased by 5 and thus christened the method of analysis. (That increment of 5 served to avoid working with negative numbers.)

Despite these firm and wide foundations, Berkson and Finney are the names most commonly associated with the logit and probit analysis, respectively. Berkson (1944, 1949) advocated the use of the logistic transformation in place of the normal (thus, 'logit' refers to 'logistic probability unit') and introduced the computationally efficient minimum chi-square estimation procedure. In 1947, Finney published the first edition of the treatise *Probit Analysis*, which became the standard reference and computational handbook for applications of the probit technique.

An application of probit to the problem of automobile demand by Farrell (1954) appears to have been its first use in economics and was inspired by the literature in bioassay. Farrell's 'stimulus' variable was income and conceptually he sought an estimate of the mean of the random income threshold above which a household would make a purchase. Soon after, Tobin (1955) examined the demand for durable goods with a probit model. He included two exogenous variables, conceived of a latent index made up of a linear combination of them as the stimulus, and framed the problem in a multiple regression analysis setting rather than the anova structure more common in bioassay. While Tobin's paper apparently did not impress contemporary editors, its wide citation in the econometrics literature over the next two decades suggests that it was indeed a landmark contribution.

Sporadic contributions to the probit–logit literature appeared throughout the 1960s. Zellner and Lee (1965) adapted Berkson's MIN chi-square estimator to multiple dichotomous relationships, Goldberger (1964) was the first to mention probit in an econometrics text, and, in work which appears to have been independent of the bioassay literature, Theil (1969) suggested a multinomial logit model and derived the weighted least squares estimator for it.

Parallel to but independent of the development of probit and logit for empirical work in bioassay, the probit and logit functions were being used in theoretical models of behaviour in psychology, and this literature led to independent introduction of the techniques to economics. Following the early work of Fechner, Urban and others, Thurstone (1927) obtained the probit model as a solution to the derivation of choice probabilities in a theoretical model of random utility, and Luce (1959) derived the multinomial logit model from an axiomatic approach to this same problem. This literature was introduced to economics by Marschak (1960) and inspired empirical research using the multinomial logit model in the area of travel demand. The new results and careful elucidation of the theoretical and statistical foundations

of the multinomial logit model by McFadden (1974; 1981) serves as a major stimulus to the application and further development of these techniques in economics.

Limited dependent variable models have a much shorter history, and, aside from univariate censored and truncated variable models, a history which is more closely confined to economics. The seminal contribution was by Tobin (1958) who considered the zero mass point and continuous positive observations on durable goods expenditures to reflect a hybridization of probit models and regression models. The truncated variable model he proposed was christened 'Tobit' by Goldberger (1964). Aside from sporadic applications and extensions through the 1960s, the next major contributions were proofs of the asymptotic properties of the MLE by Amemiya (1973) and Heckman's (1974) careful derivation of a limited variable model from an underlying theory of economic behaviour. Subsequent developments have flourished, particularly in new models and applications involving self-selection and in properties and performance of various estimators.

See Also

- ▶ Censored Data Models
- ▶ Discrete Choice Models
- ▶ Limited Dependent Variables
- ▶ Selection Bias and Self-selection

Bibliography

Amemiya, T. 1973. Regression analysis when the dependent variable is truncated normal. *Econometrica* 41: 997–1016.

Amemiya, T. 1981. Qualitative response models: A survey. *Journal of Economic Literature* 19: 1483–1536.

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

Arabmazar, A., and P. Schmidt. 1981. Further evidence on the robustness of the Tobit estimator to heteroscedasticity. *Journal of Econometrics* 17: 253–258.

Berkson, J. 1944. Application of the logistic function to bio-assay. *Journal of the American Statistical Association* 39: 357–365.

Berkson, J. 1949. Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association* 44: 273–278.

Bishop, Y., S. Fienberg, and P. Holland. 1975. *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Bliss, C. 1934. The method of probits. *Science* 79: 38–39.

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

Farrell, M. 1954. The demand for motor cars in the United States. *Journal of the Royal Statistical Society, Series A* 117: 171–193.

Fechner, G. 1860. *Elemente der Psychophysik*. Leipzig: Breitkopf & Härtel.

Finney, D. 1947. *Probit analysis*. Cambridge: Cambridge University Press.

Gaddum, J. 1933. *Reports on biological standards III. Methods of biological assay depending on a quantal response*, Special report series medical research council, vol. 183. London: H.M. Stationery Office.

Goldberger, A. 1964. *Econometric theory*. New York: Wiley.

Goldberger, A. 1983. Abnormal selection bias. In *Studies in econometrics: Time series and multivariate statistics*, ed. S. Karlin, T. Amemiya, and L.A. Goodman. New York: Academic.

Hausman, J., and D. Wise. 1977. Social experimentation, truncated distributions and efficient estimation. *Econometrica* 45: 319–339.

Heckman, J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–693.

Hurd, M. 1979. Estimation in truncated samples when there is heteroscedasticity. *Journal of Econometrics* 11: 247–258.

Lee, L.F. 1981. Simultaneous equations models with discrete and censored variables. In *Structural analysis of discrete data with econometric applications*, ed. C. Manski and D. McFadden. Cambridge, MA: Harvard University Press.

Luce, R. 1959. *Individual choice behavior: A theoretical analysis*. New York: Wiley.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic.

McFadden, D. 1981. Econometric models of probabilistic choice. In *Structural analysis of discrete data with econometric applications*, ed. C. Manski and D. McFadden. Cambridge, MA: Harvard University Press.

McKelvey, R., and W. Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4: 103–120.

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

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

Marschak, J. 1960. Binary-choice constraints and random utility indicators. In *Mathematical methods in the social sciences*, ed. K. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Nelson, F. 1981. A test for misspecification in the censored-normal model. *Econometrica* 49: 1317–1329.

Olsen, R. 1978. A note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica* 46: 1211–1215.
Powell, J. 1984. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25: 303–325.
Quandt, R. 1982. Econometric disequilibrium models. *Econometric Reviews* 1: 1–63.
Robinson, P. 1982. On the asymptotic properties of estimators of models containing limited dependent variables. *Econometrica* 50: 27–41.
Rosett, R. 1959. A statistical model of friction in economics. *Econometrica* 27: 263–267.
Rosett, R., and F.D. Nelson. 1975. Estimation of a two-limit probit regression model. *Econometrica* 43: 141–146.
Theil, H. 1969. A multinomial extension of the linear logit model. *International Economic Review* 10: 251–259.
Thurstone, L. 1927. A law of comparative judgement. *Psychological Review* 34: 273–286.
Tobin, J. 1955. *The application of the multivariate probit analysis to economic survey data*, Cowles foundation discussion paper, No. 1. New Haven.
Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.
Urban, F. 1909. Die Psychophysichen Massmethoden als Grundlagen Empirischer Messungen. *Archiv für die Gesammte Psychologie* 15: 261–355.
Zellner, A., and T. Lee. 1965. Joint estimation of relationships involving discrete random variables. *Econometrica* 33: 383–394.

Lognormal Distribution

P. E. Hart

Abstract

If there is a number, Θ, such that $Y = \log_e(X - \Theta)$ is normally distributed, the distribution of X is *lognormal.* The important special case of $\Theta = 0$ gives the two parameter lognormal distribution, $X \sim \Lambda(\mu, \sigma^2)$ with $Y \sim N(\mu, \sigma^2)$, where μ and σ^2 denote the mean and variance of $\log_e X$. The classic work on the subject is by Aitchison and Brown (1957). A useful survey is provided by Johnson et al. (1994, ch. 14). They also summarize the history of this distribution: the pioneer contributions by Galton (1879) on its genesis, and by McAlister (1879) on its measures of location and dispersion, were followed by Kapteyn (1903), who studied its genesis in more detail and also devised an analogue machine to generate it. Gibrat's (1931) study of economic size distributions was a most important development because of his law of proportionate effect. Since then there has been an immense number of applications of the lognormal distribution in the natural, behavioural and social sciences.

Keywords

Central limit theorems; Gibrat's Law; Lognormal distribution

JEL Classifications

C1

If there is a number, Θ, such that $Y = \log_e(X - \Theta)$ is normally distributed, the distribution of X is *lognormal.* The important special case of $\Theta = 0$ gives the two parameter lognormal distribution, $X \sim \Lambda(\mu, \sigma^2)$ with $Y \sim N(\mu, \sigma^2)$, where μ and σ^2 denote the mean and variance of $\log_e X$. The classic work on the subject is by Aitchison and Brown (1957). A useful survey is provided by Johnson et al. (1994, ch. 14). They also summarize the history of this distribution: the pioneer contributions by Galton (1879) on its genesis, and by McAlister (1879) on its measures of location and dispersion, were followed by Kapteyn (1903), who studied its genesis in more detail and also devised an analogue machine to generate it. Gibrat's (1931) study of economic size distributions was a most important development because of his law of proportionate effect. Since then there has been an immense number of applications of the lognormal distribution in the natural, behavioural and social sciences.

Why does the lognormal distribution appear to occur so frequently? One plausible answer is based on the central limit theorems used to explain the genesis of a normal curve. If a large number of random shocks, some positive, some negative, change the size of a particular variable, X, in an additive fashion, the distribution of that variable will tend to become normal as the number of shocks increases. But if these

shocks act multiplicatively, changing the value of X by randomly distributed proportions instead of absolute amounts, the central limit theorems apply to $Y = \log_e X$ which tends to be normally distributed. Hence X has a lognormal distribution.

The substitution of multiplicative for additive random shocks generates a positively skew, leptokurtic, lognormal distribution instead of the symmetric, mesokurtic normal curve. But the degree of skewness and kurtosis of the two-parameter lognormal curve depends solely on σ^2, so if this is low enough, the lognormal approximates the normal curve. The important difference is that X cannot take zero or negative values which may make the lognormal distribution a more appropriate representation of variables, such as height and weight, which must take positive values. Clearly, the widespread occurrence of positive variables in practice, coupled with the great flexibility of the shape of the lognormal, provide further reasons for its frequent application.

See Also

- ▶ Gini Ratio
- ▶ Inequality (Measurement)
- ▶ Lorenz Curve
- ▶ Pareto Distribution

Bibliography

Aitchison, J., and J.A.C. Brown. 1957. *The lognormal distribution*. Cambridge: Cambridge University Press.

Galton, F. 1879. The geometric mean in vital and social statistics. *Proceedings of the Royal Society of London* 29: 365–367.

Galton, F. 1889. *Natural inheritance*. London: Macmillan.

Galton, F. 1892. *Hereditary genius*. London: Macmillan.

Gibrat, R. 1931. *Les inégalites économiques*. Paris: Libraire du Recueil Sirey.

Johnson, N., S. Kotz, and L. Balakrishnan. 1994. *Continuous Univariate Distributions*. Vol. 1. New York: John Wiley.

Kapteyn, J.C. 1903. *Skew frequency curves in biology and statistics*. Astronomical Laboratory, Groningen: Noordhoff.

McAlister, D. 1879. The law of the geometric mean. *Proceedings of the Royal Society of London* 29: 367–375.

Long Cycles

John Cornwall

For over a century many economists have maintained that the historical development of capitalist economies can be usefully described in terms of regular, re-occurring cyclical movements in overall economic activity. Depending upon the investigator, these regularities have been revealed allegedly in re-occurring 3–5 year inventory or Kitchin cycles, 7–11 year investment or Juglar cycles, 15–25 year building or Kuznets cycles and 45–60 year long-wave or Kondratieff cycles.

From the beginning of the postwar period until the early 1970s, interest in the study of these cycles was slight. The long postwar boom in the capitalist world shifted the interests of the economics profession from the study of cycles, their identification and causes, to the determinants of growth and differences in growth rates between countries and over time.

The collapse of the worldwide boom in the early 1970s lead to a marked renewal of interest in discovering patterns of cyclical behaviour, this time with a special interest in long-term movements of output and its rate of growth. This is seen in the explanation given by long-wave theorists for the economic difficulties of the 1970s and 1980s. The slowdown in growth rates in the second half of the 1960s and the beginning of a period of low and even zero growth in the 1970s and 1980s were interpreted as a reflection of the recession and depression phases, respectively, of the fourth long cycle experienced in the capitalist world. Moreover, according to some long-wave theorists the present recession and stagnation must await the accumulation of some very basic structural and institutional changes before recovery can set in motion a fifth long cycle.

Early Work

In spite of Marx's own rejection of any renewal possibilities under capitalism, early long-wave

writers tended to be Marxists (e.g. van Gelderen 1913; Kondratieff 1926; Parvus 1901; de Wolff 1924). In these writings the underlying causal mechanism generating the long cycle was poorly spelt out. The explanation of cyclical turning points was even less satisfactory. For example, rather than explain turning points as part of the endogenous mechanism that generated the cycle, exogenous factors had to be introduced in an *ad hoc* manner.

The statistical support for all the early versions of the theory was extremely weak. Often data on output of selected industries – for example, pig iron – and even price data were used as an indicators of movements in the overall activity of the economy. Furthermore, given the desire to explain cycles of 45–60 years duration and the relatively short period over which data had been collected, verification was made difficult by the small sample of observations of possible long waves.

Schumpeter

Since an important characteristic of growth historically has been the changing composition of output and the industrial structure, many long-cycle theorists came to view the long wave as something related to innovations in production processes and the introduction of new consumer goods. Schumpeter's work on the long cycle (1939) reflects this view and marks what can be considered the second stage in the evolution of long-cycle theory. Innovational investment and its diffusion play a key role in his theory in determining both the amplitude and the duration of the long cycle.

Following the depression phase of the previous long cycle, a period that Schumpeter referred to as a period of 'creative destruction', certain conditions for recovery accumulated. Unfortunately neither the list of necessary conditions for recovery nor the length of time required for their accumulation was specified. The recovery was revealed first in the rising profitability of investment followed by an investment spurt dominated by innovational investment in new industries. As the boom progressed, the rapid growth of output in these industries provided the basis for a prolonged and strong boom throughout the economy as the investment boom spread throughout the economy.

Eventually, however, the growth of the capital stock so expanded the capacity of the economy relative to the growth of demand that the profitability of further investment was greatly reduced, both in the older industries and those new sectors that dominated the boom. The recession and eventually the depression phases of the long cycle then followed.

The Postwar Phase

The long-wave tradition of analysis was kept alive in the early postwar period by writers such as Mandel (1964). But it is useful to think of a third phase in the evolution of long-cycle analysis that begins with the worldwide stagnation and stagflation of the early 1970s (Freeman 1982; Mensch 1975; van Duijn 1983). In this phase greater attention and detail is given to different types of innovations, for example, process versus product innovations, cost-reducing versus expansionary innovations; and greater diversity in the assumed timing of the different kinds of innovations and the inventions leading to innovations is noticeable.

For example, it is argued that innovations are clustered in the depression phase of the long cycle because they have been crowded out in the previous boom phases of the same cycle (Mensch 1975). In contrast, other long-wave theorists see basic process innovations clustered in the boom phase of the cycle, when capacity is being strained and expansionary investment projects are needed (Freeman 1982). Still other writers emphasize the importance of basic scientific discoveries in determining the timing of both inventions and innovations (Rosenberg 1974).

In contrast to long-cycle theories that trace their origins to Schumpeter, some postwar studies deny the importance of innovations as the dominant force behind the long wave. The expansion and contraction of production in the capital goods

industry interacting with lags has been stressed (Forrester 1977). A modern Marxist interpretation emphasizes fluctuations in the rate of profit as the primary driving mechanism (Mandel 1978). Eclectic theories that combine these various explanations have also been developed (van Duijn 1983).

The 'Inevitability' Issue

A tendency in the new long-wave theories to downgrade a major role for policy in altering the shape of the long cycle is also apparent in the most recent versions of the theory. Renewal depends upon an accumulation of a backlog of inventions and innovations and not upon, say, stimulative aggregate demand policies which might be thought to reduce macro risks, improve profit prospects and thereby stimulate investment as excess capacity is reduced.

Instead, long-wave advocates often argue that stimulating aggregate demand through higher government expenditures only leads to a further deterioration of economic conditions (van Duijn 1983) since this 'crowds out' investment including the innovative kind. As a result the policy implications of long-wave theory often suggest waiting for the inevitable recovery following either an extended period during which the capital stock depreciates or becomes obsolescent or some kind of drastic social and economic reorganization (Mandel 1964).

The policy measures that are acceptable to long-wave economists include a reduction in the size of the welfare state and a decrease in taxes for the wealthy in hopes of stimulating savings and eventually investment (van Duijn 1983). Modern-day long-cycle theory in some versions thus provides a justification for what has come to be known as supply-side economics and a strong rejection of Keynesian-type aggregate-demand policies. It must be stressed, however, that the existence of the long wave or cycle is anything but proven. Neither the statistical evidence nor the theoretical arguments have persuaded more than a small minority of economists of their existence.

See Also

- ▶ Kondratieff Cycles
- ▶ Kuznets Swings
- ▶ Long Swings in Economic Growth

Bibliography

De Wolff, S. 1924. Prosperitäts- und Depressionsperioden. In *Der lebendige Marxismus: Festgabe zum 70 Geburstage von Karl Kautsky*, ed. O. Jensen. Jena: Thuringer Verlagsanstalt.

Forrester, J. 1977. Growth cycles. *De Economist* 125(4): 525–543.

Freeman, C. 1982. *The economics of industrial innovations*, 2nd ed. Cambridge, MA: MIT.

Kondratieff, N. 1926. Die langen Wellen der Konjunktur. *Archiv fur Sozialwissenschaft und Sozialpolitik* 56(3): 573–609.

Mandel, E. 1964. The economics of neo-capitalism. *Socialist Register* 1: 56–67.

Mensch, G. 1975. *Das technologische Patt*. Frankfurt an Main: Umschau Verlag.

Parvus [Alexander Helphand]. 1901. *Die Handelskrisis und die Gewerkschaften*. Munich: Verlag M. Ernst.

Rosenberg, N. 1974. Science, invention and economic growth. *Economic Journal* 84(333), March, 90–108.

Schumpeter, J. 1939. *Business cycles*. 2 vols, New York: McGraw-Hill.

Van Duijn, J. 1983. *The long wave in economic life*. London: George Allen & Unwin.

Van Gelderen, J. [Fedder, J.] 1913, April–June. Springvloed. *De Nieuwe Tijd 18*, 253–277; 369–384; 445–464.

Long Memory Models

P. M. Robinson

Abstract

Time series exhibiting varying forms of strong dependence are considered. Stationary parametric and semiparametric models, and their estimation, are first discussed. We go on to review nonlinear, nonstationary and multivariate models.

Keywords

ARMA processes; Cointegration; Fourier frequencies; Fractional autoregressive integrated

moving average (FARIMA); Fractional noise; Generalized method of moments (GMM); Long memory models; Maximum likelihood; Multivariate models; Nonlinear models; Non-stationary models; Semiparametric estimation; Statistical inference; Time series analysis; Whittle estimates

JEL Classification
C1

Much analysis of economic and financial time series focuses on stochastic modelling. Deterministic sequences, based on polynomials and dummy variables, can explain some trending or cyclic behaviour, but residuals typically exhibit serial dependence. Stochastic components have often been modelled by stationary, weakly dependent processes: parametric models include stationary and invertible autoregressive moving average (ARMA) processes, while a non-parametric approach usually focuses on a smooth spectral density. In many cases, however, we need to allow for a greater degree of persistence or 'memory'. This is characterized by stationary time series whose autocorrelations are not summable or whose spectral densities are unbounded, or by non-stationary series evolving over time. The latter are partly covered by unit root processes, but considerably greater flexibility is possible.

Basic Models

Early empirical evidence of slowly decaying autocorrelations emerged long ago, in analyses of astronomical, chemical, agricultural and hydrological data, and then in economics and finance. A stationary parametric model which attracted early interest is 'fractional noise'. Let x_t, $t = 0, \pm 1, \ldots$, be a covariance stationary discrete time process, so its autocovariance $cov(x_t, x_{t+u})$ depends only on u, and thus may be denoted by γ_u. Then fractional noise x_t has autocovariance

$$\gamma_u = \gamma_0 \left\{ |u+1|^{2d+1} - 2|u|^{2d+1} + |u-1| \right\}^{2d+1}, \quad u = 0, \pm 1, \ldots, \tag{1}$$

where the parameter d is called the 'memory parameter', and satisfies $-\frac{1}{2} < d < \frac{1}{2}$. When $d = 0$ (1) implies that $\gamma_u = 0$ for $u \neq 0$, so x_t is white noise. But if $0 < d < \frac{1}{2}$, we have

$$\gamma_u \sim 2d\left(d + \frac{1}{2}\right)\gamma_0 |u| 2^{d-1}, \quad \text{as } |u| \to \infty, \tag{2}$$

where '$\sim$' means that the ratio of left- and right-hand sides tends to one. It follows from (2) that γ_u does decrease with lag u, but so slowly that

$$\sum_{u=-\infty}^{\infty} \gamma_u = \infty. \tag{3}$$

In the frequency domain, when x_t has a spectral density $f(\lambda)$, $\lambda \in (-\pi, \pi)$ given by

$$f(\lambda) = (2\pi)^{-1} \sum_{\infty=-\infty}^{\infty} \gamma_u \cos(u\lambda), \lambda \in (-\pi, \pi),$$

the property (3) is equivalent to

$$f(0) = \infty, \tag{4}$$

and more precisely a fractional noise process x_t has spectral density satisfying

$$f(\lambda) \sim C\lambda^{-2d}, \text{as } \lambda = 0+. \tag{5}$$

In general we can regard (3) and (4) as basic indicators of a 'long memory' process x_t, and (2) and (5) as providing more detailed description of autocorrelation structure at long lags, or spectral behaviour at low frequencies. By contrast, if x_t were a stationary ARMA, γ_u would decay exponentially and $f(\lambda)$ would be analytic at all frequencies. The structure (5) is similar to Granger's (1966) 'typical spectral shape of an economic variable'.

The model (1) is connected with the physical property of 'self-similarity', and, so far as

economic and financial data are concerned, found early application in work of Mandelbrot (1972) and others. However, (1) imposes a very rigid structure, with autocorrelations decaying monotonically and depending on a single parameter. In addition, though a formula for $f(\lambda)$ corresponding to (1) can be written down, it is complicated, and (1) does not connect well mathematically with other important time series models, and does not lend itself readily to forecasting.

An alternative class of 'fractionally integrated' processes leads to a satisfactory resolution of these concerns. This is conveniently expressed in terms of the lag operator L, where $Lx_t = x_{t-1}$. Given the formal expansion

$$(1-s) = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} s^j,$$

we consider generating x_t from a zero-mean stationary sequence u_t, $t = 0; \pm 1; \ldots$, by

$$(1-L)^d (x_t - \mu) = v_t, \tag{6}$$

where $\mu = Ex_t$ and

$$|d| < \frac{1}{2}.$$

If v_t has absolutely summable autocorrelations, that satisfy some mild additional conditions, both the properties (2) and (5) hold. In the simplest case of (6), v_t is a white noise sequence. Then γ_u decays monotonically when $d \in \left(0, \frac{1}{2}\right)$ and indeed behaves very much like (1). This model may have originated in Adenstedt (1974), though he stressed the case $d \in \left(-\frac{1}{2}, 0\right)$, where x_t is said to have 'negative dependence' or 'antipersistence'. Taking v_t to be a stationary and invertible ARMA process, with autoregressive order p and moving average order q, gives the FARIMA (p, d, q) process of Granger and Joyeux (1980). In principle, the short memory process v_t in (6) can be specified in any number of ways so as to yield (2) and/or (5); a process satisfying this condition is sometimes called $I(d)$.

Statistical Inference

Given observations x_t, $t = 1; \ldots; n$ there is interest in estimating d. If v_t has parametric autocorrelation, as when x_t is a FARIMA (p, d, q), one can form a Gaussian maximum likelihood estimate of d and any other parameters. This estimate has the classical properties of being $n^{1/2}$ – consistent and asymptotically normal and efficient. Computationally somewhat more convenient estimates, called Whittle estimates, have the same asymptotic properties. Indeed, for standard FARIMA (p, d, q) parameterizations, say, the estimates of d and of ARMA coefficients have asymptotic variance matrix that is unaffected by many departures from Gaussianity. Though these asymptotic properties are of the same type as one obtains for estimates of short memory processes, such as ARMAs, their proof is considerably more difficult (see Fox and Taqqu 1986), due to the spectral singularity (4). In econometrics, generalized method of moments (GMM) estimation has become very popular, and GMM estimates have been proposed for long memory models. However, unless a suitable weighting is used, they are not efficient under Gaussianity, are not more robust asymptotically to non-Gaussianity, and are not even asymptotically normal when $d > \frac{1}{4}$.

If the parametric autocorrelation is mis-specified, for example if in the FARIMA (p, d, q) p or q are chosen too small or both are chosen too large, then the procedures described in the previous paragraph will generally produce inconsistent estimates of d, as well as of other parameters. Essentially, the attempt to model the short memory component of x_t damages estimation of the long memory component. This difficulty can be tackled by a 'semiparametric' approach, if one regards the local or asymptotic specifications (2) or (5) as the model, and estimates d using only information in low frequencies or in long lags. Frequency domain versions are by far the more popular here, having the nicest asymptotic statistical properties. In the log periodogram estimate of d, logged periodograms are regressed on a logged local approximation to $f(\lambda)$, over the m Fourier frequencies closest to the origin

(Geweke and Porter-Hudak 1983), m having the character of a bandwidth number similar to those used in smoothed nonparametric functional estimation. An alternative approach optimizes a local Whittle function, again based on the lowest m Fourier frequencies (Künsch 1987). In the asymptotics for both types of estimate (see Robinson 1995a, b) m must increase with n, but more slowly (to avoid bias); both the log periodogram and local Whittle estimates are $m^{1/2}$ – consistent and asymptotically normal, with the latter the more efficient (though it is computationally more onerous, being only implicitly defined). Because both converge more slowly than estimates of correctly specified parametric models, a larger amount of data may be necessary for estimates to be reasonably precise. Moreover, estimates are sensitive to the choice of However, automatic and other rules are available for determining m; and semiparametric methods of estimating memory parameters have become very popular not only because of the robust character of the asymptotic results, but because of their relative simplicity.

The long memory processes we have been discussing exhibit an excess of low frequency power (5). But one can also consider parametric or semiparametric models for a spectral density with one or more poles at non-zero frequencies. These models can be used to describe seasonal or cyclic behaviour (see Arteche and Robinson 2000). It is also possible to estimate the unknown location of a pole, that is, cycle (see Giraitis et al. 2001).

Nonlinear Models

In non-Gaussian series, not all information is contained in first and second moments. In particular, in many financial series observations x_t may appear to have little or no autocorrelation, but instantaneous nonlinear functions, such as squares x_t^2, exhibit long memory behaviour. We can develop models to describe such phenomena. For example, let

$$x_t = \varepsilon_t h_t, \tag{7}$$

where x_t is a sequence of independent and identically distributed random variables with unit variance, whereas h_t is a stationary autocorrelated sequence, such that ε_s and h_t are independent for all s, t. Then for all $u \neq 0$, $cov = (x; x_{t+u}) = 0$ but $cov\left(x_t^2, x_{t+u}^2\right) = cov\left(h_t^2, h_{t+u}^2\right)$, which in general can be non-zero. In particular, if h_t^2 has long memory, so has x_t^2. In a more fundamental modelling we can take h_t to be a nonlinear function of an underlying long memory Gaussian processes, with the functional form of h determining the extent of any long memory in h^2; these issues were discussed in some generality by Robinson (2001). The models form a class of long memory stochastic volatility models, whose estimation has been discussed by Hurvich et al. (2005), for example.

The fractional class (6) can be modified or extended to describe a wide class of nonstationary behaviour. For $d \geq \frac{1}{2}$ the variance of x_t (6) explodes, but we can consider truncated versions such as

$$x_t = (1 - L)^{-d}\{v_t 1(t \geq 1)\}$$

where $1(\cdot)$ is the indicator function, or

$$x_t = (1 - L)^{-k}\{w_t 1(t \geq 1)\}$$

for integer $k \geq 1$, where w_t is a stationary $I(c)$ process, $|c| < \frac{1}{2}$, and $d = k + c$. In either case we might call x_t a (nonstationary) $I(d)$ process, for $d \geq \frac{1}{2}$. Both models include the unit root case $d = 1$ that has proved so popular in econometrics. However, the fractional class $I(d)$, for real-valued d, bridges the gap between short memory and unit root processes, allowing also for the possibility of arbitrarily long memory d. The 'smoothness' of the $I(d)$ family is associated with classical asymptotic theory, which is not found in autoregressive based models around a unit root. Robinson (1994) showed that Lagrange multiplier tests for the value of d, and any other parameters, have asymptotic null χ^2 distributions for all real d. Also, under nonstationary suitably modified parametric and semiparametric methods of estimating d, extending those for the stationary case, tend still to be respectively $n^{1/2}$- and $m^{1/2}$ – consistent, and asymptotically normal, unlike, say, the lag-one sample autocorrelation of a unit root series.

Multivariate Models

Often in economics and finance we are concerned with a vector of jointly dependent series, so x_t is vector-valued. Such series can be modelled, either parametrically or semiparametrically, to have long memory, with different elements of x_t possibly having different memory parameters, and being stationary or nonstationary. Methods of statistical inference developed for the univariate case can be extended to such settings. However, multivariate data introduces the possibility of (fractional) cointegration, where a linear combination of x_t (the cointegrating error) can have smaller memory parameter than the elements of x_t. Cointegration has been extensively developed for the case x_t is $I(1)$ and cointegrating errors are $I(0)$, and methods developed for this case can fail to detect fractional cointegration. Moreover, it is possible for stationary series, not only nonstationary ones, to be fractionally cointegrated, as seems relevant in financial series. In either case, methods of analysing cointegration that allow memory parameters of observables and cointegrating errors to be unknown (see, for example, Hualde and Robinson 2004) afford considerable flexibility.

See Also

- ▶ Central Limit Theorems
- ▶ Econometrics
- ▶ Non-parametric Structural Models
- ▶ Semiparametric Estimation
- ▶ Time Series Analysis

Research supported by ESRC Grant R000239936.

Bibliography

Adenstedt, R. 1974. On large-sample estimation for the mean of a stationary random sequence. *Annals of Statistics* 2: 1095–1107.

Arteche, J., and P. Robinson. 2000. Semiparametric inference in seasonal and cyclic long memory processes. *Journal of Time Series Analysis* 21: 1–25.

Fox, R., and M.S. Taqqu. 1986. Large sample properties of parameter estimates of strongly dependent stationary Gaussian time series. *Annals of Statistics* 14: 517–532.

Geweke, J., and S. Porter-Hudak. 1983. The estimation and application of long memory time series models. *Journal of Time Series Analysis* 4: 221–238.

Giraitis, L., J. Hidalgo, and P. Robinson. 2001. Gaussian estimation of parametric spectral density with unknown pole. *Annals of Statistics* 29: 987–1023.

Granger, C. 1966. The typical spectral shape of an economic variable. *Econometrica* 34: 150–167.

Granger, C., and R. Joyeux. 1980. An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–39.

Hualde, J., and P. Robinson. 2004. *Semiparametric estimation of fractional cointegration*. Mimeo: London School of Economics.

Hurvich, C., E. Moulines, and P. Soulier. 2005. Estimating long memory in volatility. *Econometrica* 73: 1283–1328.

Künsch, H. 1987. Statistical aspects of self-similar processes. In *Proceedings of the First World Congress of the Bernoulli Society*, vol. 1, ed. Y. Prohorov and V. Sazonov. Utrecht: VNU Science Press.

Mandelbrot, D. 1972. Statistical methodology for non-periodic cycles: From the covariance to R/S analysis. *Annals of Economic and Social Measurement* 1: 259–290.

Robinson, P. 1994. Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association* 89: 1420–1437.

Robinson, P. 1995a. Log-periodogram regression of time series with long range dependence. *Annals of Statistics* 23: 1048–1072.

Robinson, P. 1995b. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* 5: 1630–1661.

Robinson, P. 2001. The memory of stochastic volatility models. *Journal of Econometrics* 101: 192–218.

Long Run and Short Run

Carlo Panico and Fabio Petri

Abstract

The notion of long-run and short-run equilibrium was introduced by Marshall in 1890 and reflected the ‘long-period method’ of analysis in use among classical political economists since the 18th century. In the early 1930s, dissatisfaction with some of the neoclassical conclusions led to a shift to different methods of analysis and to the introduction of new equilibrium notions. These changes, together with

the tendency to use old terminology for new equilibrium concepts, have deprived the terms 'short-period' and 'long-period' of a uniform meaning and have been a source of confusion and misunderstandings in recent debates on theoretical and applied work.

Keywords

Average interest rate; Capital endowment; Circulating capital; Diminishing marginal returns; Effective demand; Fixed capital; General equilibrium; Intertemporal equilibria; Keynes, J. M.; Long-period method; Long-run and short-run; Market price; Marshall, A.; Marx, K. H.; Natural price; Natural rate and market rate of interest; Partial equilibrium; Secular equilibrium; Short-run and long-run equilibrium; Stationary equilibria; Stationary state; Supply and demand; Temporary equilibrium; Underemployment equilibrium; Value and distribution, neoclassical theory of

JEL Classifications

C32

The distinction between long-run and short-run (or long-period and short-period) equilibrium, introduced by Marshall (see Marshall 1890, pp. 363–80; hints at this distinction are also to be found in some of Marshall's early works, dated 1870–71, recently re-presented in Whitaker 1975, pp. 119–64), reflected a method which was the generally accepted one at the time, and essentially the same as the method of the classical political economists and of Marx. The use of the method was not affected by the deep change undergone by the theory of value and distribution around the 1870s with the advent of what is nowadays called the 'neoclassical' school. This method, called 'method of long-period positions' (Garegnani 1976), however, has been abandoned in much of the modern mainstream work on value. Further, there is no uniform meaning attributed to the terms 'short-period' and 'long-period', but rather a variety of usages depending on the theoretical framework of the writer, a situation responsible for many misunderstandings and debates at cross purposes.

The Classical Political Economists

Since its origin in the writings of 18th-century authors, economic theory has used what has been subsequently named the 'long-period method' of analysis to investigate how production, distribution and accumulation take place within a market economy. According to Quesnay and A. Smith, the system 'market economy' produces results which are 'independent of men's will' (Quesnay 1758). Competition, Smith thought, tends to establish uniformity in the 'average' or 'natural' rates of wages, profits and rent. 'Market' prices, that is, observed prices, thus tend to gravitate towards their 'natural' levels (also called 'average prices' or 'prices of production'), defined as those which allowed the payment of wages, profits and rents at their average or natural rates (Smith 1776, pp. 57–61).

According to the classical political economists, a divergence between the 'market' and the 'natural' price of a commodity is caused by a divergence between the amount supplied by producers and the 'effectual demand' for it, that is, '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, p. 58). This divergence implies windfall profits or losses for that commodity. If supply coincides with 'effectual demand', 'market' price corresponds to 'natural' price. The rate of profit earned in that sector is equal to the one which is uniformly earned in the whole economy. Equilibrium conditions are said to prevail. Within this approach, therefore, fluctuations of supply and demand explain nothing but the deviations of 'market' prices from 'natural' prices.

The idea that the interaction of competitive market forces pushes the actual level of economic variables towards their 'natural' or 'average' level was applied to different fields of economic theory. Marx, for instance, applied it to the analysis of the 'market' and the 'average' interest rate (see Marx 1972, pp. 355–66). The latter rate, according to Marx, was determined by 'the average conditions of competition, the balance between lender and borrower' (Marx 1972, p. 363) in the money

market over a certain historical period (Marx 1972, p. 363). He rejected previous views determining this rate in terms of 'natural' laws, like the rate of growth of timber in central Europe forests (Marx 1972, p. 363 n.) or in terms of the rate of return on capital invested in the productive sectors depending upon the material or technological conditions of production of commodities (Marx 1972, p. 363). In his historically relative determination, the 'average' interest rate, being constrained by no 'natural' or 'material' law, can be at any level. At the same time, the interaction of demand and supply determines the daily variations of the 'market' interest rate and makes it converge towards its 'average' level.

The application of the 'long-period method' to the analysis of the interest rate makes it clear that the essential element of the method is the reference to an 'average' or 'normal' position around which the actual values of the variable considered gravitate. Reference to the attainment of a uniform rate of profit in all sectors is not strictly necessary if the theory does not determine the variable considered on the basis of the technological conditions of production. In Marx's analysis, since the 'average' interest rate is independent of the rate of profits, it is possible to separate the study of the factors determining the former rate from the study of the technological links between distributive variables and commodity prices, where competitive forces set in motion a gravitation process when windfall profits or losses appear in particular industries. The notion of 'average' interest rate, which may be used to identify a position of long-period equilibrium for this variable, can thus be introduced and analysed by referring to a normal position of this variable, which has actually prevailed over a certain period, without making reference to a uniform rate of profits. In a theory determining the 'natural' interest rate on the basis of technological conditions of production, instead, no separation can be made between the analysis of the average interest rate and that of the links between commodity prices and distributive variables. In this case, the condition of a uniform rate for return on capital defines the 'long-period equilibrium' position for both commodity prices and interest rate.

The Rise of Neoclassical Economics

The long-period method was also used by those economists (like Walras, Menger, Jevons, Böhm-Bawerk, J.B. Clark, Wicksell, et al.) who some years later introduced and developed the 'neoclassical' theory of value and distribution. *No question was raised by these authors as to the use of this method*.

The new theory, unlike the previous one, determined prices, output and distribution simultaneously. The 'natural' or 'equilibrium' values of all these variables (including the interest rate and the level of activity in the economy, which turns out to be a full employment level) depended, among other things, upon the technological conditions of production and were thus associated with the attainment of a uniform rate of profits in the economy.

Among the earlier neoclassical economists, Marshall deserves special consideration, since he introduced the notion of short- and long-period equilibrium (see Marshall 1890, p. 80). In his writings, Marshall tried to show how the neoclassical principles of price determination in terms of supply and demand functions could be applied to analytical levels which were closer to actual events. He thus analysed price determination for each single market (partial equilibrium) and within this analysis he referred to three different notions of equilibrium (temporary, short-period and long-period), which differed as to the conditions determining the supply functions. In a temporary equilibrium, it was supposed, there is no time to change the supply of the commodity. The amount supplied is fixed and the equilibrium price is that which allows that quantity to be demanded.

Analyses of short-period equilibrium assume that there is time to change supply through production, but there is no time to change the structure of *fixed capital* goods existing in *that* industry. This assumption constrains the technological possibilities of production. As in the case of temporary equilibrium, short-period equilibrium is compatible with windfall profits or losses.

In long-period analyses, it is assumed instead that there is time to adapt the structure of fixed capital goods of the industry so that quasi-rents

(that is, entrepreneurial net profits) disappear. The price then guarantees just the 'normal rate of profits' (that is, the 'equilibrium' real rate of return on capital which is uniform in the whole economy).

Marshall's partial equilibrium analysis appears to rely on general equilibrium analysis for the determination of the 'equilibrium' rate of return on capital and of 'ceteris paribus' prices. The view that the 'general equilibrium' analysis was logically prior appears accepted in some major contributions of the debate on Marshall's theory of value of the 1920s and the early 1930s (see Sraffa 1925 and 1926, and Pigou's reply, 1927). Marshall's starting point thus was the same as that of Walras, Wicksell, and of the other neoclassical economists mentioned above.

Long-period general equilibrium must not be confused with 'secular' equilibrium, which results from allowing enough time for factor endowments to change under the influences of demographic factors and propensity to save, so as to cause the economy to reach 'stationary' or 'steady growth' conditions (see Robbins 1930).

Short- and Long-Period in Keynes

By the end of the 1920s, dissatisfaction with the neoclassical conclusions as to the level of activity of the economy and with the analysis of capital led some economists to new analytical developments, which affected for the first time the method used too.

J.M. Keynes criticized the neoclassical conclusion that the market economy has an inherent tendency towards full employment. In the preparatory works and in the introduction to the *General Theory* he insisted that his concern was not the analysis of the temporary and cyclical fluctuations of the level of activity, but the theory dealing with the more fundamental forces which tend to prevail in the economic system (see Keynes 1936, pp. 4–5, 1973, pp. 405–7, and 1979, pp. 54–7). He wanted thus to replace the neoclassical long-period theory of the level of output with a new one. Yet the way he presented his new theory has raised many problems of interpretation also related to the method used.

First of all, Keynes stated in his book that he assumed as given the structure of *fixed capital* goods existing in the economy. This can lead to consider his theory as a short-period one, arguing that it would determine the level of capacity utilization in the economy. It is difficult, however, to support this interpretation also with the argument that in the *General Theory* Keynes was following Marshall's definition of short-period, which was confined to partial equilibrium analysis. Marshall knew that the time required for adjustment of the structure of fixed capital goods differed from one industry to the other, so that it would have been unreasonable to extend the hypothesis of a fixed structure from one industry to the whole economy, as Keynes did. This element of ambiguity as to the use of the concepts has raised many puzzling questions among the interpreters of Keynes.

At the same time, Keynes explicitly stated that his theory was meant to explain why the level of employment, over a specific historical period, oscillates round an intermediate or average position (often not a full-employment one), whereas in other periods it oscillates round a different one (Keynes 1936, p. 254). This reference to 'specific historical periods' and to 'average or normal positions' can lead to consider Keynes's theory as a long-period one, in the same way as Marx's theory of the 'average' interest rate. The assumption of a fixed structure of capital goods would thus play a secondary role in Keynes's theory.

Besides, Keynes hinted towards an analysis of accumulation which emphasizes the role played by effective demand (Keynes 1936, pp. 372–80). The trend followed by a growing economy in which adjustment in the structure of fixed capital goods has occurred, is affected by the level of effective demand. The possibility of assuming in this analysis an adjusted structure of fixed capital goods (to which a uniform rate of profits corresponds) can lead to consider this as the *long-period theory* present in the *General Theory.*

Finally, the maintenance in the *General Theory* of elements belonging to the neoclassical tradition, like the acceptance of the principle of diminishing marginal returns for capital from which the existence of a full-employment level of the rate of interest is derived (see Keynes

1936, pp. 147–8, 178, 203, 235 and 243, 1973, pp. 456, 615, 630) has allowed some interpreters to consider Keynes's 'underemployment equilibrium' as a situation in which market forces have not yet worked out their effects fully, consequently defining it as a position of 'short-period equilibrium' (see Patinkin 1976, pp. 116–19; Winch 1969, p. 167).

The presence of several lines of development of its basic principle (that of effective demand) and the lack of precision and coherence as to the concepts and the analytical elements used appear to be an endless source of discussion as to the interpretation of Keynes's work. The existing evidence does not seem to support, however, the view that the *General Theory* wanted to move along the same lines as Hayek, Hicks and others, who in those years were proposing the neoclassical theory of value, distribution and the level of output on the basis of a method of analysis different from the long-period one.

Post-Walrasian Developments

In the same years, dissatisfaction with the neoclassical analysis of capital was leading to a shift in method, owing to the adoption of what may be called 'post-Walrasian' notions of general equilibrium, elaborated by Hayek and Lindahl around the 1930s, but first proposed to a wider audience in 1939 by Hicks's *Value and Capital* (see Garegnani 1976; Milgate 1979). The change in method derives from the change in the treatment of the capital endowment.

In the traditional neoclassical treatment, dominant up to the 1950s, the conception of equilibrium as a centre of gravitation of time-consuming adjustments (a conception incompatible with taking as given the equilibrium endowments of the several capital goods) had been reconciled with the supply-and-demand approach to factor pricing by conceiving capital as a *single* factor of production, capable of changing 'form' (that is, of embodying itself into different vectors of heterogeneous capital goods) without changing in 'quantity', so that its 'form' (that is, composition) could be left to be determined by the equilibrium condition of a uniform rate of return on the supply price of capital goods – the distinguishing element of long- period positions. Capital so conceived had ultimately to be measured as an amount of value, because in equilibrium different capital goods earn rewards proportional to their values. Within the neoclassical framework, therefore, the reference to a homogeneous factor ' capital', a value magnitude, was a logical necessity, entailed by the attempt to explain distribution through the equilibrium between demand for and supply of 'factors of production', without abandoning the traditional method of longperiod positions (Petri 2004). With one exception, this conception of capital was in fact more or less explicitly adopted by all founders of neoclassical theory and it was the target of the Cambridge critique of the 1960s (Harcourt 1969; Garegnani 1970). The only exception had been Walras, who intended as well to determine a long-period equilibrium and accordingly maintained the uniform-profit-rate condition, but took as data the endowment of each kind of capital goods, with the result that his model was generally devoid of solutions.

Walras's treatment of the capital endowment as a given vector is maintained in post-Walrasian general equilibrium analyses, but the condition of uniform profit rate on supply price is dropped. Existing capital goods are treated like natural resources; commodities are dated, so prices of future commodities are distinguished from prices of currently available commodities; and the current composition of the production of new capital goods is determined in either of two ways: by assuming the existence of complete futures markets (intertemporal equilibria, see for example Debreu 1959), or through the introduction of expectations among the data (temporary equilibria, see for example Hicks 1939 and Grandmont 1977).

The difference between the notion of equilibrium entailed by such a treatment of capital and that entailed by the long-period method of analysis warrants emphasis (Garegnani 1990). The latter attempts to represent states of the economy which have the role of centres of gravitation of observed day-to-day magnitudes: chance movements away from such a state set off forces

tending to bring the economy back to it. Changes in the economy can then be studied by comparing the long-period positions corresponding to the situation before and after the change. Post-Walrasian equilibria cannot have such a role, because they rely on data some of which (the endowments of capital goods and, where futures markets are not complete, expectations) would be altered by any chance deviation from the equilibrium: thus the forces set off by this deviation would not tend to bring the economy back to the same equilibrium. For the same reason, stability questions relative to post-Walrasian equilibria can only be asked for imaginary atemporal adjustment processes which exclude the implementation of disequilibrium production decisions before the equilibrium is reached.

A Variety of Usages

The introduction of new equilibrium concepts, together with the tendency to overlook the existence of differences with previous ones and to use the same terminology for the former and for the latter, has been a source of confusion and misunderstandings in recent debates on theoretical and applied work.

The term 'short-period equilibria' has been sometimes applied to post-Walrasian equilibria (including 'fix price' equilibria with quantity adjustments, which share the same impermanence of data). On other occasions, Keynes's notion of equilibrium has been identified with temporary equilibrium. In both cases, the very great difference between Marshall's and Keynes's analyses on one side and post-Walrasian analyses on the other side has been neglected: in post-Walrasian models, *all* capital goods, including circulating capital goods, are given, while in Marshall's short period analyses only the fixed plant of a single industry is a datum, and in Keynes's work only the fixed capital goods of the whole economy are given.

At the same time, the term ' long-period equilibrium' has been used in recent years to refer (a) to post-Walrasian intertemporal equilibria with futures markets extending far into the future; (b) to sequences of temporary equilibria; (c) to stationary or steady-growth equilibria. In all these cases, an incomplete grasp of the changes introduced in the notion of equilibrium appears to emerge.

Finally, modern neoclassical economists sometimes develop applied analyses using the traditional method of long-period positions, although rejecting, as their theoretical foundations, the traditional versions of neoclassical theory in favour of the post-Walrasian ones, which are not compatible with that method.

See Also

▶ Marshall, Alfred (1842–1924)

Bibliography

Debreu, G. 1959. *Theory of value*. New York: Wiley.
Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37: 407–436.
Garegnani, P. 1976. On a change in the notion of equilibrium in recent work on value and distribution. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. Amsterdam: North-Holland.
Garegnani, P. 1990. Quantity of capital. In *Capital theory*, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.
Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45: 535–572.
Harcourt, G.C. 1969. Some Cambridge controversies in the theory of capital. *Journal of Economic Literature* 7: 369–405.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Keynes, J.M. 1936. The general theory of employment, interest and money. In *The collected writings of J.M. Keynes*, ed. D. Moggridge, vol. VII. London: Macmillan.
Keynes, J.M. 1973. The general theory and after. Part I: Preparation. In *The collected writings of J.M. Keynes*, ed. D. Moggridge, vol. XIII. London: Macmillan.
Keynes, J.M. 1979. The general theory and after: A supplement. In *The collected writings of J.M. Keynes*, ed. D. Moggridge, vol. XXIX. London: Macmillan.
Marshall, A. 1890. Principles of economics. 8th ed. London: Macmillan, 1920.
Marx, K. 1972. *Capital*. Vol. 3. London: Lawrence &Wishart.
Milgate, M. 1979. On the origin of the notion of 'intertemporal equilibrium'. *Economica* 46: 1–10.

Patinkin, D. 1976. *Keynes' monetary thought*. Durham: Duke University Press.
Petri, F. 2004. *General equilibrium, capital and macroeconomics. A key to recent controversies in equilibrium theory*. Cheltenham: Edward Elgar.
Pigou, A.C. 1927. The laws of diminishing and increasing cost. *Economic Journal* 37: 188–197.
Quesnay, F. 1758. *Tableau économique.* Reproduced and trans. as *Quesnay's Tableau économique,* ed. M. Kuczynski and R.L. Meek. London: Macmillan, 1972.
Robbins, L. 1930. On a certain ambiguity in the conception of stationary equilibrium. *Economic Journal* 40: 194–214.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations.* Ed. E. Cannan. London: Methuen, 1904.
Sraffa, P. 1925. Sulle relazioni fra costo e quantità prodotta. *Annali di Economia* 2: 277–328.
Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36: 535–550.
Whitaker, J.K., ed. 1975. *The early economic writings of Alfred Marshall, 1867–1890*. London: Macmillan.
Winch, D. 1969. *Economics and policy: A historical study.* London: Hodder & Stoughton.

Long Swings in Economic Growth

C. Freeman

The notion of half-century long swings in the growth of industrial capitalism was advanced by several economists, including Pareto, already before World War I, on the basis of empirical observations of long-term trends in the 19th-century statistics of prices, interest rates and trade. It was, however, a Dutch Marxist, Van Gelderen (1913) who was the first to attempt a systematic description and explanation of the phenomenon. He used the expressions 'spring tide' and 'ebb-tide' to describe the alternating periods of buoyant expansion and relative stagnation which he claimed to detect.

Nikolai Kondratiev, by whose name long cycles or swings in economic development are most widely known, was apparently unaware of Van Gelderen's earlier work (in Dutch), when he published his own empirical and theoretical studies in the 1920s (Kondratiev 1925). Although he and Schumpeter (1939) used the expression 'long cycles', other authors have preferred the term 'long waves' (e.g. Mandel 1972) or 'long swings', but essentially they are all discussing the same phenomenon.

While he was Director of an Institute of Applied Economic Research in Moscow in the 1920s, Kondratiev sought to demonstrate the existence of long cycles from the 1770s onwards, on the basis of historical data for the leading industrial countries. At the time and ever since, this historical statistical evidence has been disputed, both by his more orthodox Marxist critics in the USSR and by economists outside (e.g. Weinstock 1964). Some of the main points at issue have been the statistical techniques (moving averages and trend analysis) and the limitations of the statistical data, particularly for the earlier periods when the only production series was for specific commodities. Time series for investment and employment were even more deficient. Of the original 25 time series which Kondratiev (1925) cited in support of his theory the majority covered two cycles and only three of them covered all three cycles.

The criticism of Kondratiev's ideas and of other long wave theories has of course never been confined to the issue of the statistical evidence. Already in the 1920s Kondratiev's Soviet critics pointed to the problem of exogenous 'shocks' to the world economy, such as wars, revolutions and gold discoveries and their effects on long-term fluctuations in prices and production. They also pointed to the variety of national circumstances in the duration and intensity of cyclical crises and booms and to the *new* features associated with each successive historical period. Finally, they disputed Kondratiev's attempt to explain long waves in terms of the replacement cycle for very long-lived types of (mainly infrastructural) fixed investments.

It is highly unlikely that those who believe that long swings in economic life are a significant phenomenon, which merits some serious research and explanation, will ever satisfy their statistical critics, if only for the obvious reason that four cycles is still far too small a number on which to base any firm generalizations. Some historians associated with the journal *Annales* have claimed

that the long swings go back to the Middle Ages. They base their claims largely on data relating to agricultural prices and among the possible explanations of long swings in the price of grain are theories of climatic fluctuations related to sun-spots (Braudel 1979).

The contemporary debate on long waves relates, however, almost entirely to the fluctuations in the development of industrialized economies over the last two centuries. Whilst the critics of Kondratiev believed that they had disposed of his theories and discredited his statistical techniques, it is hardly surprising that the deep depression of the 1930s and the depressed conditions of the 1980s both gave rise to renewed interest in long wave theories. Indeed the course of economic development in the 20th century appears to follow a long wave pattern much more consistently and obviously than that of the 19th century. If one of the tests of a theory is predictive power, then it must be said that Kondratiev's analysis advanced in the 1920s gave a rather reliable forecast of the main trends in the world economy over the next sixty years.

However, most advocates of long wave theories deny deterministic explanations with fixed periodicity, maintaining only that the deeper structural crises, which have affected some, if not all, capitalist economies at intervals of approximately half a century, merit some special attention and explanation. They also point out that general equilibrium theory has not been particularly helpful in understanding the fluctuations in the long-term growth process of the world economy, or of national economies.

The growing interest in the problems of long-term fluctuations was not confined to economists of any particular 'school'. Among those who have written books on the subject or contributed papers in the recent debate were both neoclassical (Glismann et al. 1980) and Marxist (Mandel 1980) economists, and some, such as Rostow (1978), who could be described as broadly Keynesian in their approach, or Dupriez (1947), who might be described as a monetary theorist. One striking feature of the international conferences on long waves (IIASA 1983 and 1985) was the evident revival of interest in the subject on the part of Russian and other East European economists and their renewed attempt to explain long waves in orthodox Marxist terms (Kuczynski 1985). Another group actively involved was the Systems Dynamics group at MIT led by Forrester (1981) who developed a long-term dynamic model of the US economy, which displays long wave characteristics based on alternating periods of over- or under-investment in the capital goods sector.

At the heart of the long wave debate in the 1980s has been the Schumpeterian interpretation of Kondratiev's long cycles. Indeed many economists became aware of 'Kondratiev cycles' largely because of Schumpeter's adoption of the idea and his attempt to provide an explanation in terms of successive waves of technical innovations or, as he described it, 'creative gales of destruction'. Schumpeter suggested that each of the major upswings in the economy, which Kondratiev had detected, was based on a wave of new investment associated with the spread of one or several major new technologies, such as steam power and electric power. Schumpeter maintained that the growth process in capitalist societies was not simply *accompanied* by technical and organizational innovations, but was *driven* by such innovations. Since he believed that innovations were spread unevenly over time and across different sectors of the economy, it was consistent to regard them as the main source of disequilibrium in the system and as the source of a variety of cyclical fluctuations, including long cycles in the case of major new technologies.

The critics of a Schumpeterian explanation of long waves follow Kuznets (1940) in questioning whether any innovations could be so great in their impact on the economy as to cause major fluctuations in investment behaviour and the economy more generally. Among the most interesting and influential attempts to provide a plausible explanation of this relationship were those of Mensch (1975) and of Perez (1983, 1985).

Mensch suggested that radical innovations were bunched together during periods of deep

depression, such as the 1830s, the 1880s and the 1930s. He explained this bunching in terms of the pressures on entrepreneurs to adopt novel solutions which they were unwilling to risk during boom periods when things were going well. Freeman et al. (1982) disputed the empirical evidence on the bunching of innovations as well as the theoretical explanation and suggested that the *diffusion* of clusters of interrelated innovations ('new technological systems') was more important in understanding cyclical fluctuations than the dates of discrete radical innovations.

Carlota Perez (1983, 1985) criticized Schumpeter for his failure to develop a satisfactory theory of deep depressions. She pointed out that although he offered a plausible explanation of investment booms in terms of the rapid diffusion of new technologies and the associated 'bandwagon' and 'swarming behaviour' of entrepreneurs, and could also explain recessions in terms of the 'competing away' of profit margins during diffusion and the limits to growth of any particular technology, he regarded the deeper depressions as 'pathological' and was unconvincing in his attempts to explain why newly emerging technologies should not take over the expansionary momentum. Kuznets (1940) had also spotted this weakness in Schumpeter's theory of long cycles and asked ironically whether the heroic entrepreneurs got tired every 50 years.

Perez pointed out that in considering the introduction of revolutionary new technologies into the economic system, it is necessary to take into account the institutional and social framework as well as the economic sub-system more narrowly conceived. The really big changes in technology, such as the contemporary introduction of computerized information technology or the earlier introduction of energy-intensive mass and flow production systems, or of electric power, inevitably entail big changes in social institutions, as well as changes in company organization, patterns of investment and the skill profile of the work-force. But whereas technology changes very rapidly, there is considerable inertia in social institutions, as well as resistance from group interests associated with older technologies, sectors of the economy and occupations, who may feel their very existence is threatened by revolutionary changes in technology. Consequently for Perez depressions are the symptom of a structural mis-match between the potential of an emerging new paradigm in technology and a socio-institutional framework which is geared to an older (but now obsolescent) technological paradigm. Only when there have been far-reaching changes in social institutions and ways of organizing business can the full productivity potential of the new technology be realized. It follows from this mode of conceptualizing long swings in economic growth that the radical innovations which crystallize in a 'new technology system' or in a new 'techno-economic paradigm' have in many cases been introduced already *before* a period of depression (contrary to Mensch's theory) but their widespread diffusion in many branches of the economy is hampered or prevented by the mismatch in social institutions, skills and capital stock (Freeman and Perez 1986).

L

See Also

- ▶ Kondratieff Cycles
- ▶ Long Cycles

Bibliography

Braudel, F. 1979. *Civilisation matérielle, économie et capitalisme, XVe-XVIIIe siècle*, vol. 3. Paris: Colin.

Delbeke, J. 1981. Recent long wave theories: A critical survey. *Futures* 13(4): 246–257.

Dupriez, L.H. 1947. *Des mouvements économiques généraux*. Louvain: Institut de recherches économiques et sociales.

Forrester, J.K. 1981. Innovation and economic change. *Futures* 13(4): 323–331.

Freeman, C. (ed.). 1984. *Long waves in the world economy*. London: Frances Pinter.

Freeman, C., and C. Perez. 1986. *The diffusion of technical innovations and changes of techno-economic paradigm*. DAEST, Venice Conference.

Freeman, C., J. Clark, and L.L.G. Soete. 1982. *Unemployment and technical innovation: A study of long waves in the world economy*. London: Frances Pinter.

Glismann, H.H., et al. 1980. *Lange Wellen Wirtschaftlichen Wachstums*, Kiel Discussion Paper, vol. 74. Kiel: Institut fur Weltwirtschaft.

IIASA (International Institute of Applied Statistical Analysis). 1983 and 1985. *Reports of international conferences on long waves*. Vienna: IIASA.
Kondratiev, N. 1925. The major economic cycles. *Voprosy Konjunktury* 1: 28–79. English trans. in *Review of Economic Statistics* 17, November 1935, 105–15. Reprinted in *Lloyd's Bank Review* 129, July 1978, 41–60.
Kuczynski, T. 1985. *Marx and Engels on long waves*. Paper contributed to IIASA Conference, Weimar.
Kuznets, S. 1940. Schumpeter's business cycles. *American Economic Review* 30: 257–271.
Mandel, E. 1972. *Der Spätkapitalismus*. Frankfurt am Main: Surkampf. Revised trans. as *Late Capitalism*, London: New Left Books, 1975.
Mandel, E. 1980. *Long waves of capitalist development: The Marxist interpretation*. Cambridge: Cambridge University Press.
Mensch, G. 1975. *Das technologische Patt: Innovationen überwinden die Depression*. Frankfurt: Umschau. Trans. as Stalemate in Technology: Innovations Overcome the Depression, New York: Ballinger.
Pareto, V. 1913. Alcuni relazioni fra la stato, sociale e la variazioni della prosperita economica. *Revista Italiana di Sociologica*: 5011–548.
Perez, C. 1983. Structural changes and the assimilation of new technology in the economic and social system. *Futures* 15(4): 357–375.
Perez, C. 1985. Micro-electronics, long waves and world structural change: New perspectives for developing countries. *World Development*, Special Issue, 13: 441–63.
Rostow, W.W. 1978. *The world economy: History and prospect*. London: Macmillan.
Schumpeter, J.A. 1939. Business cycles: A theoretical, historical and statistical analysis of the capitalist process, 2 vols. New York: McGraw-Hill.
Van Duijn, J.J. 1983. *The long wave in economic life*. London: Allen & Unwin.
Van Gelderen J. (as J. Fedder.) 1913. Springvloed, beschovwingen over industriele ontwikkeling en prijsbeweging. *De Nieuwe Tijd* 18: 253–77; 369–84; 445–64.
Weinstock, W. 1964. *Das Problem der Kondratioff-Zyklen*. Berlin: Duncker & Humblot.

Longe, Francis David (1831–c1905)

A. Picchio

Longe graduated from Oriel College, Oxford, in 1854 and joined the Bar in 1858. He was associated with the Children's Employment Committee, served as Inspector of the Local Government Board, and was private secretary to Lord Goschen, president, between 1868 and 1870, of the Poor Law Board (Hollander 1903, p. 3).

Longe's main contribution to the history of economic analysis is his original, lucid and direct attack against J.S. Mill's formulation of the theory of the wages fund. His critical assessment predated Thornton's attempts in the same direction in 1867 and 1869, which led to speculations about possible plagiarism (Hollander 1903; Schumpeter 1954, pp. 669–70).

Longe attacked the view that wages were determined by a quantitative relationship between a given fund (aggregate demand for labour) and population (aggregate supply). The critique involved methodology, political perspectives, the notion of supply and demand price– quantity relationships, the nature of the aggregates, the definition of labour, and most of all the notion of a determinacy of wages as equilibrium price. His approach to the labour market derives from the classical tradition of a 'natural' wage based on a 'customary standard, which however much it may be ignored by theorists, is the immediate basis on which the wages or remuneration of every trade and profession rests' (Longe, 1866, p. 16). His acquaintance as a barrister with the institutional and conflictual aspects of the labour market – expressed also in his historical work on strikes, praised by the Webbs ((Webb and Webb 1894) 1920, pp. 227–8) – also contributed to his view of the labour market. Longe's approach is still relevant to the modern debate on the labour-market structure, and it is quite superior to Thornton's work *On Labour* both in analytical and stylistic terms, although in the form of a pamphlet.

Longe confutes the wages fund theory and shows the non-existence of any definite or mechanical relationship between some supposed given amount of capital and a definite number of labourers. The capital applicable for the payment of wages is not distinct from general wealth and there is no definite fund which is 'destined' for the purchase of labour. The notion of full employment, obtainable by downward flexible wages, was questioned also

because the supply of labour is formed by an heterogenous dependent population which changes its structure according to habits and customs and does not react normally to changes in wages and competition. Longe's critique relates not only to the 'vulgar' assumption of a definite wages fund (Marshall 1975, p. 818) – to which J.S. Mill easily responded – but to the whole methodology of supply-and-demand determined wages. His work could have induced a greater caution to the introduction of new theories of distribution based on endogenously determined wages.

Selected Works

1860. An inquiry into the law of strikes. Cambridge/London (Goldsmith Library). Reprinted in National Association for the Promotion of Social Science, *Report on Trade Societies and Strikes*, 1860.
1866. *A Refutation on the Wage-fund theory of modern political economy as enunciated by Mr Mill, M.P and Mr Fawcett, M.P.* London: Longmans, Green.
1867. The law of trade combinations in France. *Fortnightly Review.*
1883. *A critical examination of Mr George's 'Progress and Poverty' and Mr Mill's theory of wages*. London: Simpkin & Marshall.

Bibliography

Foster, J. 1885. Men at the bar. A biographical handilist of the members of the various Inns of Court. London.
Hollander, J.H. 1903. Introduction to F.D. Longe. In *The wages fund theory.* Baltimore: Johns Hopkins Press.
Marshall, A. 1975. *The early economic writings of Alfred Marshall, 1867–1890*, Vols I and II. ed. J.K. Whitaker, London: Macmillan.
McNulty, P.J. 1980. *The origins and development of labor economics*. Cambridge, MA: Harvard University Press.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Webb, S., and B. Webb. 1894. *The history of trade unionism*. London: Longman & Co. 2nd revised edn, 1920.

Longfield, Mountifort (1802–1884)

R. D. Collison Black

JEL Classifications
B31

Longfield was born at Desertserges, Country Cork, Ireland, in 1802. Although he graduated from Trinity College, Dublin, in 1823 with first class honours in natural sciences, he was elected a Fellow of his college in 1825 as 'jurist'. His subsequent career was primarily in real property law, but when Archbishop Whately founded the professorship of political economy at Trinity College, Dublin, in 1832, Longfield was the successful candidate and became the first holder of the chair, from 1832 until 1836. In 1834 he was appointed Regius Professor of Feudal and English Law and in 1849 became one of the first Commissioners of the newly established Irish Incubered Estates Commission. When this was transmuted into the Landed Estates Court in 1858, Longfield was appointed a Judge of that court, retiring in 1867. He died in Dublin in 1884.

In 1847 he was one of the founder members of the Dublin Statistical Society (later re-named the Statistical and Social Inquiry Society of Ireland) and followed Whately as its President in 1863, but his many other public services derived primarily from his positions as advocate and judge. In his later years Longfield never returned to political economy but continued to write on questions of Irish land tenure and social reform.

The three volumes of lectures which Longfield published during his tenure of the Whately chair attracted little attention at the time, but have since been recognized as containing contributions to economic theory of outstanding originality. In his *Lectures on Political Economy* (1834a) Longfield dealt with the central issues of classical theory, those of value and distribution, in a manner which displayed a very clear grasp of the structure of Ricardian theory, but which in content

diverged fundamentally from Ricardo's approach. He laid stress on the determination of market rather than natural values and presented remarkably complete demand-and- supply theory supplemented by elements of utility analysis. Perhaps his most original contribution was made in the area of distribution, where he formulated a theory of profits as determined by the marginal productivity of physical capital and a theory of wages as determined by the specific productivity of the labourer.

Longfield rejected the idea that the 'natural price' of labour was determined by subsistence, arguing that the 'wages of the labourer depend upon the value of his labour and not upon his wants' (1834a, p. 206). Although, like Ricardo, Longfield predicted a rise in rents, a fall in profits and a rise in wages in the progress of society, his view of the long-term prospects for economic growth was optimistic. He expected the effects of increased population to be offset by technical progress in agriculture, and foresaw many benefits from the increased accumulations of capital which would lower profits, not least among them the increased productivity of labour, which would raise wages.

Longfield's two other published courses of lectures are more concerned with current economic problems, but his *Lectures on Commerce* (1835) contained several anticipations of later developments in international trade theory. His analysis of the causes of international specialization extended to all variations in factor endowments and he specifically treated the case of trade in more than two commodities, showing that each country would tend to export those commodities in which the productivity of its labour was above average and import those in which it was below average.

In his *Lectures on Poor Laws* (1834b) Longfield endorsed Senior's stern principle that assistance to the able-bodied should be confined to the barest subsistence – perhaps, ironically, because of the very optimism of his views about the likely trends of profits and wages. On the other hand, he favoured generous public assistance to those unable, through age or disability, to fend for themselves – even to the extent of advocating non-contributory old-age pensions. Longfield repeated this proposal in 1872, when he specifically considered state interference with the distribution of wealth; unlike most of his contemporaries he was then prepared also to advocate public dispensaries and hospitals to which access would not be means-tested, improved sanitary regulation of housing standards, free public education and improved public recreation facilities.

Longfield's economic writings appear to have had little influence on his contemporaries, but since his rediscovery by Seligman (1903) the originality of his contributions has come to be generally recognized.

Selected Works

1834a. *Lectures on political economy, delivered in Trinity and Michaelmas terms, 1833*. Dublin: R. Milliken & Son. Reprinted, 1971.

1834b. *Four lectures on poor laws delivered in Trinity term, 1834*. Dublin: R. Milliken & Son. Reprinted, 1971.

1835. *Three lectures on commerce, and one on absenteeism, delivered in Michaelmas term, 1834*. Dublin: William Curry, Junior & Co. Reprinted, 1971.

1840. Banking and currency. *Dublin University Magazine* 15: 3–15; 218–233; 369–389, 609–620. Reprinted, 1971.

1870. Tenure of land in Ireland. In *Systems of land tenure*. London: Cobden Club.

1872a. The limits of state interference with the distribution of wealth in applying taxation to the assistance of the public. *Journal of the Statistical and Social Inquiry Society of Ireland* 6:105–114.

1872b. *Elementary treatise on series*. Dublin.

1971. *The economic writings of Mountifort Longfield*, ed. R.D. Collison Black. New York: A.M. Kelley.

Bibliography

Black, R.D.C. 1945. Trinity College, Dublin, and the theory of value 1832–1863. *Economica*, NS, 12: 140–148.

Black, R.D.C. 1984. The Irish dissenters and nineteenth-century political economy. In *Economists and the Irish*

economy, ed. A.E. Murphy, 120–137. Dublin: Irish Academic Press.
Moss, L.S. 1976. *Mountifort Longfield, Ireland's first professor of political economy*. Ottowa: Green Hill Publishers.
Murphy, A.E. 1984. Mountifort Longfield's appointment to the chair of political economy in trinity College, Dublin, 1832. In *Economists and the Irish economy*, ed. A.-E. Murphy, 13–24. Dublin: Irish Academic Press.
Seligman, E.R.A. 1903. On some neglected British economists. *Economic Journal* 13: 335–363; 511–535. Revised version published in *Essays in economics,* ed. E.-R.A. Seligman. New York: Macmillan, 1925.

Longitudinal Data Analysis

Cheng Hsiao

Abstract

The advantages and fundamental methodological issues of statistical inference using data sets that contain time series observations of a number of individuals are discussed.

Keywords

Central limit theorems; Discrete choice models; Generalized method of moments; Instrumental variables; Laws of large numbers; Least squares; Linear models; Logit models; Maximum likelihood; Panel data; Tobit models; Unit roots

JEL Classifications

C23; C33

Why Panel Data?

'Longitudinal data' (or 'panel data') refers to datasets that contain time series observations of a number of individuals. In other words, it provides multiple observations for each individual in the sample. Compared with cross-sectional data, in which observations for a number of individuals are available only for a given time, or time-series data, in which a single entity is observed over time, panel data have the obvious advantages of more degrees of freedom and less collinearity among explanatory variables, and so provide the possibility of obtaining more accurate parameter estimates. More importantly, by blending inter-individual differences with intra-individual dynamics, panel data allow the investigation of more complicated behavioural hypotheses than those that can be addressed using cross-sectional or time-series data.

For instance, suppose a cross-sectional sample yields an average labour participation rate of 50 per cent for married women. Given that the standard assumption for the analysis of cross-sectional data is that, conditional on certain variables, each woman is a random draw from a homogeneous population, this would imply that each woman has a 50 per cent chance of being in the labour force at any given time. Hence, a married woman would be expected to spend half of her married life in the labour force and half out of it. The job turnover would be frequent, and the expected average job duration would be just two years (Ben-Porath 1973). However, the cross-sectional data could be drawn from a heterogeneous population in which 50 per cent of the sample was drawn from the population that always works and 50 per cent from the population that never works. In this situation, there is no turnover and a woman's current work status is a perfect predictor of her future work status. To discriminate between these two possibilities, we need information on individual labour-force histories in different sub-intervals of the life cycle, which can be provided only if information is available on the intertemporal dynamics of individual entities. On the other hand, although time series data provide information on dynamic adjustment, variables over time tend to move collinearly, hence making it difficult to identify micro-dynamic or macro-dynamic effects. Often, estimation of distributed lag models has to rely on strong prior restrictions like the Koyck or Almon lag, with very little empirical justification (for example, Griliches 1967). With panel data, the inter-individual differences can often lessen the problem of multicollinearity and provide the

possibility of estimating unrestricted time adjustment patterns (for example, Pakes and Griliches 1984).

By utilizing information on both the intertemporal dynamics and the individuality of the entities, panel data may also allow an investigator to control the effects of missing or unobserved variables. For instance, MaCurdy's (1981) life-cycle labour supply of prime-age males with perfect foresight model assumes that the logarithm of hours worked is a linear function of the real wage rate and the logarithm of the worker's marginal utility of initial wealth, which is unobserved. Since the wage rate and the marginal utility of initial wealth are correlated, any instrument that is correlated with the wage rate will be correlated with the marginal utility of initial wealth. There is no way one can obtain a consistent estimate of the coefficient of the wage rate with cross-sectional data. But, if panel data are available and since marginal utility of initial wealth stays constant over time, one can take the difference of the labour supply model over time to get rid of the marginal utility of initial wealth as an explanatory variable. Regressing change in hour on change in wage rate and other socio-demographic variables can yield consistent estimates of the coefficient of the wage rate and other explanatory variables.

Panel data may also provide microfoundations for aggregate data analysis. Aggregate data analysis often invokes the 'representative agent' assumption. If micro units are heterogeneous, the time series properties of aggregate data may be very different from those of disaggregate data (for example, Granger 1990; Lewbel 1994) and policy evaluation based on aggregate data could also be grossly misleading (for example, Hsiao et al. 2005). By providing time series observations for a number of individuals, panel data are ideal for the investigation of the homogeneity issue.

Panel data involve observations of two or more dimensions. In normal circumstances, one would expect the computation and inference of panel data models to be more complicated than those of cross-section or time series data. However, in certain situations the availability of panel data actually simplifies inference. For instance, statistical inference for non-stationary panel data can be complicated (for example, Phillips 1986). But, if observations are independently distributed across cross-sectional units, central limit theorems applied across cross-sectional units lead to asymptotically normally distributed statistics (for example, Levin et al. 2002; Im et al. 2003).

Issues of Panel Data Analysis

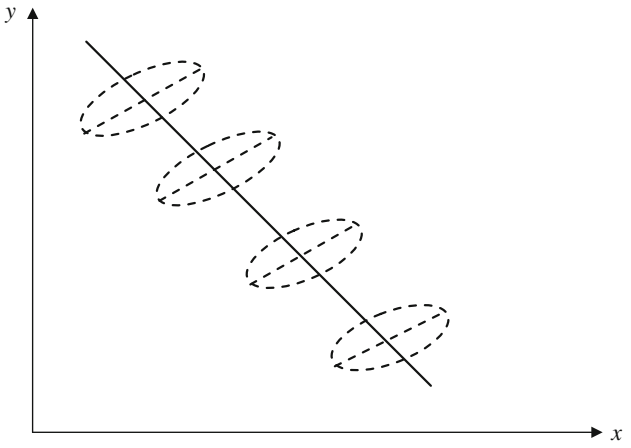
Standard statistical methodology is based on the assumption that the outcomes, say y, conditional on certain variables, say $\underset{\sim}{x}$, are random outcomes from a probability distribution that is characterized by a fixed dimensional parameter vector, $\theta, f\ (y|\ \underset{\sim}{x};\ \underset{\sim}{\theta}\ )$. For instance, the standard linear regression model assumes that $f(y|\underset{\sim}{x};\underset{\sim}{\theta})$ takes the form that $E(y|\underset{\sim}{x}) = \alpha + \underset{\sim}{\beta}'\underset{\sim}{x}$, and $\mathrm{Var}(y|\underset{\sim}{x}) = \sigma^2$, where $\underset{\sim}{\theta}' = (\alpha, \underset{\sim}{\beta}', \sigma^2)$. Panel data, by their nature, focus on individual outcomes. Factors affecting individual outcomes are numerous. It is rare to be able to assume a common conditional probability density function of y conditional on $\underset{\sim}{x}$ for all crosssectional units, i, at all time, t. If the conditional density of y given $\underset{\sim}{x}$ varies across i and over t, the fundamental theorems for statistical inference, the laws of large numbers and central limit theorems, will be difficult to implement. Ignoring the heterogeneity across i and over t that are not captured by $\underset{\sim}{x}$ can lead to severely biased inference. For instance, suppose that the data is generated by

$$y_{it} = \alpha_i + \underset{\sim}{\beta}'\underset{\sim}{x}_{it} + v_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{1}$$

as depicted by Fig. 1 in which the broken-time ellipses represent the point scatter of individual observation around the mean, represented by the broken straight lines. If an investigator ignores the presence of unobserved individual-specific effects, α_i, and mistakenly estimates a model of the form

$$y_{it} = \alpha + \underset{\sim}{\beta}'x_{it} + v^*_{it} \tag{2}$$

the following equation solid line in Fig. 1 would depict the pooled least squares regression result



Longitudinal Data Analysis, Fig. 1 Scatter diagram of (y_{it}, x_{it})

which could completely contradict the individual relation between y and $\underset{\sim}{x}$.

One way to restore homogeneity across i and/or over t is to add more conditional variables, say $\underset{\sim}{z}$,

$$f\left(y_{it}|\underset{\sim}{x}_{it},\underset{\sim}{z}_{it};\underset{\sim}{\theta}\right). \tag{3}$$

However, the dimension of $\underset{\sim}{z}$ can be large. A model is a simplification of reality, not an exact representation of reality. The inclusion of $\underset{\sim}{z}$ may confuse the fundamental relationship between y and x, in particular when there is a shortage of degrees of freedom or multicollinearity, and so on. Moreover, $\underset{\sim}{z}$ may not be observable. If an investigator is interested only in the relationship between y and $\underset{\sim}{x}$, one approach to characterize the heterogeneity not captured by $\underset{\sim}{x}$ is to assume that the parameter vector varies across i and over t, $\underset{\sim}{\theta}_{it}$, so that the conditional density of y given $\underset{\sim}{x}$ takes the form $f\left(y_{it}|\underset{\sim}{x}_{it};\underset{\sim}{\theta}_{it}\right)$. However, without a structure being imposed on $\underset{\sim}{\theta}_{it}$, such a model has only descriptive value; it is not possible to draw any inference on $\underset{\sim}{\theta}_{it}$ from observed data.

One primary focus of methodological panel data literature is to suggest possible structures for $\underset{\sim}{\theta}_{it}$. One way to impose some structure on $\underset{\sim}{\theta}_{it}$ is to decompose $\underset{\sim}{\theta}_{it}$ into $\left(\underset{\sim}{\beta},\underset{\sim}{\gamma}_{it}\right)$, where $\underset{\sim}{\beta}$ is the same across i and over t, referred to as *structural parameters*, and $\underset{\sim}{\gamma}_{it}$ as *incidental parameter*s because when observations in cross-sectional units and/or time series units increase, there are rising numbers of $\underset{\sim}{\gamma}_{it}$ to be estimated. The focus then will be on how to make valid inference on $\underset{\sim}{\beta}$ after controlling the impact of $\underset{\sim}{\gamma}_{it}$.

Without imposing structure for $\underset{\sim}{\gamma}_{it}$, again it is not possible to make any inference on $\underset{\sim}{\beta}$ because the unknown $\underset{\sim}{\gamma}_{it}$ will exhaust all available sample information. On the assumption that the impacts of observable variables, $\underset{\sim}{x}$, are the same across i and over t, represented by the structure parameters, $\underset{\sim}{\beta}$, the incidental parameters $\underset{\sim}{\gamma}_{it}$ represent the heterogeneity across i and over t that are not captured by $\underset{\sim}{x}_{it}$. They can be considered as composed of the effects of omitted individual time-invariant, α_i, period individual-invariant, λ_t, and individual time-varying variables, δ_{it}. The individual time-invariant variables are variables that are the same for a given crosssectional unit through time but that vary across cross-sectional units, such as individual-firm management, ability, gender, and socio-economic background. The period individual-invariant variables are variables that are the same for all crosssectional units at a given time but that vary though time, such as prices, interest rates, and widespread optimism or pessimism. The individual time-varying variables are variables that vary across cross-sectional units at a given point in time and also exhibit variations through time, such as firm profits, sales and capital stock. The unobserved heterogeneity as represented by the individual-specific effects, α_i and time specific effects, λ_t, or individual time-varying effects, δ_{it} can be assumed to be

either random variables (referred to as the *random effects* model) or fixed parameters (referred to as the *fixed effects* model).

Linear Static Models

A widely used panel data model assumes that the effects of observed explanatory variables, $\underset{\sim}{x}$, are identical across cross-sectional units, i, and over time, t, while the effects of omitted variables can be decomposed into the individual-specific effects, α_i, time-specific effects, λ_t, and individual time-varying effects, $\delta_{it} = u_{it}$, as follows:

$$y_{it} = \underset{\sim}{\beta}' \underset{\sim}{x}_{it} + \alpha_i + \lambda_t + u_{it}, \quad \begin{matrix} i = 1, ..., N, \\ t = 1, ..., T. \end{matrix} \tag{4}$$

In a single equation framework, individual time effects, u, are assumed random and uncorrelated with $\underset{\sim}{x}$, while α_i and λ_t may or may not be correlated with $\underset{\sim}{x}$. When α_i and λ_t are treated as fixed constants, they are parameters to be estimated, so whether they are correlated with $\underset{\sim}{x}$ is not an issue. On the other hand, when α_i and λ_t are treated as random, they are typically assumed to be uncorrelated with $\underset{\sim}{x}_{it}$.

For ease of exposition, we assume that there are no time-specific effects, that is, $\lambda_t = 0$ for all t and u_{it} are independently, identically distributed (i.i.d) across i and over t. Stack an individual's T time series observations of $(y_{it}, \underset{\sim}{x}'_{it})$ into a vector and a matrix, (4) may alternatively be written as

$$\underset{\sim}{y}_i = X_i \underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i, i = 1, ..., N, \tag{5}$$

where $\underset{\sim}{y}_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (\underset{\sim}{x}_{i1}, \ldots, \underset{\sim}{x}_{iT})'$, $\underset{\sim}{u}_i = (u_{i1}, \ldots, u_{iT})'$, and $\underset{\sim}{e}$ is a $T \times 1$ vector of 1's.

Let Q be a $T \times T$ matrix satisfying the condition that $Q\,\underset{\sim}{e} = \underset{\sim}{0}$. Pre-multiplying (5) by Q yields

$$Q\underset{\sim}{y}_i = QX_i \underset{\sim}{\beta} + Q\underset{\sim}{u}_i, i = 1, ..., N. \tag{6}$$

Equation (6) no longer involves α_i. The issue of whether α_i is correlated with $\underset{\sim}{x}_{it}$ or whether α_i should be treated as fixed or random is no longer relevant for (6). Moreover, since X_i is exogenous, $E\left(QX_i\underset{\sim}{u}'_i Q'\right) = QE\left(X_i\underset{\sim}{u}'_i\right)Q' = \underset{\sim}{0}$ and $EQ\underset{\sim}{u}_i\underset{\sim}{u}'_i Q' = \frac{u2}{\sigma} QQ'$. An efficient estimator of $\underset{\sim}{\beta}$ is the generalized least squares estimator (GLS),

$$\hat{\underset{\sim}{\beta}} = \left[\sum_{i=1}^{N} X'_i Q'(QQ')^{-} QX_i\right]^{-1} \left[\sum_{i=1}^{N} X'_i Q'(QQ')^{-} QY_i\right], \tag{7}$$

where $(Q'\ Q)^{-}$ denotes the Moore–Penrose generalized inverse (for example, Rao 1973).

When $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}', Q$ is idempotent. The Moore–Penrose generalized inverse of $(Q'\ Q)^{-}$ is just $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ itself. Pre-multiplying (6) by Q is equivalent to transforming (4) into a model

$$(y_{it} - \overline{y}_i) = \underset{\sim}{\beta}'\left(\underset{\sim}{x}_{it} - \overline{\underset{\sim}{x}}_i\right) + \left(u_{it} - \overline{u}_i\right), \quad \begin{matrix} 1 = 1, ..., N, \\ t = 1, ..., T. \end{matrix} \tag{8}$$

where $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \overline{\underset{\sim}{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{x}_{it}$ and $\overline{u}_i = \frac{1}{T}\sum_{t=1}^{T} \underset{\sim}{u}_{it}$. The transformation is called *covariance transformation*. The least squares estimator (LS) (or a generalized least squares estimator, GLS) of (8),

$$\hat{\underset{\sim}{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\underset{\sim}{x}_{it} - \overline{\underset{\sim}{x}}_i\right)\left(\underset{\sim}{x}_{it} - \overline{\underset{\sim}{x}}_i\right)'\right]^{-1} \left[\sum_{t=1}^{N}\sum_{t=1}^{T}\left(\underset{\sim}{x}_{it} - \overline{\underset{\sim}{x}}_i\right)\left(\underset{\sim}{y}_{it} - \overline{\underset{\sim}{y}}_i\right)'\right], \tag{9}$$

is called *covariance estimator* or *within* estimator because the estimation of $\underset{\sim}{\beta}$ only makes use of within (group) variation of $\underset{\sim}{y}_{it}$ and $\underset{\sim}{x}_{it}$ only. The covariance estimator of $\underset{\sim}{\beta}$ turns out to be also the least squares estimator of (4) when $\lambda_t = 0$. It is the best linear unbiased estimator of $\underset{\sim}{\beta}$ if α_i is treated as fixed and u_{it} is i.i.d.

If α_i is random, transforming (5) into (6) transforms T independent equations (or observations) into $(T - 1)$ independent equations, hence the

covariance estimator is not as efficient as the efficient generalized least squares estimator if $E\alpha_i \underset{\sim}{x}'_{it} = \underset{\sim}{0}'$. When α_i is independent of $\underset{\sim}{x}_{it}$ and is independently, identically distributed across i with mean $\underset{\sim}{0}$ and variance σ^2_α, the best linear unbiased estimator (BLUE) of $\underset{\sim}{\beta}$ is GLS,

$$\hat{\underset{\sim}{\beta}} = \left[\sum_{i=1}^{N} X_i' V^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' V^{-1} \underset{\sim}{y}_i\right], \quad (10)$$

where $V = \sigma^2_u I_T + \sigma^2_\alpha \underset{\sim}{e}\underset{\sim}{e}', V^{-1} = \frac{1}{\sigma^2_u}\left[I_T - \frac{\sigma^2_\alpha}{\sigma^2_u + T\sigma^2_\alpha}\underset{\sim}{e}\underset{\sim}{e}'\right]$, Let $\psi = \frac{\sigma^2_u}{\sigma^2_u + T\sigma^2_\alpha}$, the GLS is equivalent to first transforming the data by subtracting a fraction $(1 - \psi^{1/2})$ of individual means $\overline{y}_i$ and $\overline{x}_i$ from their corresponding y_{it} and $\tilde{x}_{it}$, then regressing $[y_{it} - (1 - \psi^{1/2})\overline{y}_i]$ on $[\underset{\sim}{x}_{it} - (1 - \psi^{1/2})\underset{\sim}{x}_i]$. (for detail, see Baltagi 2001; Hsiao 2003).

When α_i is treated as fixed, the covariance estimator is equivalent to applying LS to the transformed model (8). If a variable is time-invariant, like a gender dummy, $x_{kit} = x_{kis} = \overline{x}_{ki}$, the transformation eliminates the corresponding variable from the specification. Hence, the coefficients of time-invariant variables cannot be estimated. On the other hand, if α_i is random and uncorrelated with $\underset{\sim}{x}_i$, $\psi \neq 1$, the GLS can still estimate the coefficients of those time-invariant variables.

Dynamic Models

When the regressors of a linear model contains lagged dependent variables, say, of the form (for example, Balestra and Nerlove 1966)

$$\begin{aligned}\underset{\sim}{y}_i &= \underset{\sim}{y}_{i,-1}\gamma + X_i\underset{\sim}{\beta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i \\ &= Z_i\underset{\sim}{\theta} + \underset{\sim}{e}\alpha_i + \underset{\sim}{u}_i, i = 1, ..., N. \end{aligned} \quad (11)$$

where $\underset{\sim}{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $Z_i = (\underset{\sim}{y}_{i,-1}, X_i)$ and $\underset{\sim}{\theta} = \left(\gamma, \underset{\sim}{\beta}'\right)'$. For ease of notation, we assume that yi_0 are observable. Technically, we can still eliminate the individualspecific effects by pre-multiplying (11) by the transformation matrix $Q(Q\,\underset{\sim}{e} = \underset{\sim}{0})$,

$$Q\underset{\sim}{y}_i = QZ_i\underset{\sim}{\theta} + Q\underset{\sim}{u}_i. \quad (12)$$

However, because of the presence of lagged dependent variables, $EQZ_i\underset{\sim}{u}'_i Q' \neq 0$ even with the assumption that u_{it} is independently, identically distributed across i and over t. For instance, the covariance transformation matrix $Q = I_T - \frac{1}{T}\underset{\sim}{e}\underset{\sim}{e}'$ transforms (11) into the form

$$\begin{aligned}(y_{it} - y_i) = &\left(y_{i,t-1} - \overline{y}_{i,-1}\right)\gamma \\ &+ \left(\underset{\sim}{x}_{it} - \overline{\underset{\sim}{x}}_i\right)'\underset{\sim}{\beta} \\ &+ (u_{it} - \overline{u}_i), \quad \begin{matrix} i = 1, ..., N, \\ t = 1, ..., T. \end{matrix}\end{aligned} \quad (13)$$

where $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \overline{y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}$ and $\overline{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Although, $y_{i,t-1}$ and u_{it} are uncorrelated under the assumption of serial independence of u_{it}, the covariance between $\overline{y}_{i,-1}$ and u_{it} or $y_{i,t-1}$ and $\bar{u}_i$ is of order $(1/T)$ if $|\gamma| < 1$. Therefore, the covariance estimator of $\underset{\sim}{\theta}$ creates a bias of order $(1/T)$ when $N \to \infty$ (Anderson and Hsiao 1981, 1982; Nickell 1981). Since most panel data contain large N but small T, the magnitude of the bias can not be ignored (for example, with $T = 10$ and $\gamma = 0.5$, the asymptotic bias is -0.167).

When $EQZ_i\underset{\sim}{u}'_i Q' \neq \underset{\sim}{0}$, one way to obtain a consistent estimator for $\underset{\sim}{\theta}$ is to find instruments W_i that satisfy

$$EW_i\underset{\sim}{u}'_i Q' = \underset{\sim}{0}, \quad (14)$$

and

$$\text{rank}(W_i QZ_i) = k, \quad (15)$$

where k denotes the dimension of $\left(\gamma, \underset{\sim}{\beta}'\right)'$, then apply the generalized instrumental variable or generalized method of moments (GMM) estimator by minimizing the objective function

$$\left[\sum_{i=1}^{N} W_i\left(Q\underset{\sim}{y}_i - QZ_i\underset{\sim}{\theta}\right)\right]'\left[\sum_{i=1}^{N}(W_iQ\underset{\sim}{u}_i\underset{\sim}{u}_i'Q'W'_i)\right]^{-1}\left[\sum_{i=1}^{N} W_i\left(Q\underset{\sim}{y}_i - QZ'_i\underset{\sim}{\theta}\right)\right], \tag{16}$$

with respect to $\underset{\sim}{\theta}$ (for example, Arellano 2003; Ahn and Schmidt 1995; Arellano and Bond 1991; Arellano and Bover 1995). For instance, one may let Q be a $(T-1)\times T$ matrix of the form

$$D = \begin{bmatrix} -1 & 1 & 0 & . & . \\ 0 & -1 & 1 & . & . \\ 0 & . & . & . & . \\ . & . & . & -1 & 1 \end{bmatrix}, \tag{17}$$

then the transformation (12) is equivalent to taking the first difference of (11) over time to eliminate α_i for $t = 2, \ldots, T$,

$$\Delta y_{it} = \Delta y_{i,t-1}\gamma + \Delta\underset{\sim}{x}'_{it}\underset{\sim}{\beta} + \Delta u_{it}, \quad \begin{matrix} i = 1,\ldots,N, \\ t = 2,\ldots,T, \end{matrix} \tag{18}$$

where $\Delta = (1-L)$ and L denotes the lag operator, $Ly_t = y_{t-1}$. Since $\Delta u_{it} = (u_{it} - u_{i,t-1})$ is uncorrelated with $y_{i,t-j}$ for $j \geq 2$ and $\underset{\sim}{x}_{is}$, for all s, when u_{it} is independently distributed over time and $\tilde{x}_{it}$ is exogenous, one can let W_i be a $T(T-1)\left[K+\frac{1}{2}\right]\times(T-1)$ matrix of the form

$$W_i = \begin{bmatrix} \underset{\sim}{q}_{i2} & 0 & . & . \\ 0 & \underset{\sim}{q}_{i3} & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & \underset{\sim}{q}_{iT} \end{bmatrix}, \tag{19}$$

where

$$\underset{\sim}{q}_{it} = \left(y_{i0}, y_{i1},\ldots, y_{i,t-2}, \underset{\sim}{x}'_i\right), \underset{\sim}{x}'_i = \left(\underset{\sim}{x}'_{i1},\ldots,\underset{\sim}{x}'_{iT}\right)'$$

and $K = k-1$. Under the assumption that $(\underset{\sim}{y}i', \underset{\sim}{x}i')$ are independently, identically distributed across i, the Arellano and Bover (1995) GMM estimator takes the form

$$\hat{\underset{\sim}{\theta}}_{AB,GMM} = \left\{\left[\sum_{i=1}^{N}\underset{\sim}{Z}'_iD'W'_i\right]\left[\sum_{i=1}^{N}W_iAW'_i\right]^{-1}\left[\sum_{i=1}^{N}W_iDZ_i\right]\right\}^{-1} \left\{\left[\sum_{i=1}^{N}\underset{\sim}{Z}'_iD'W'_i\right]\left[\sum_{i=1}^{N}W_iAW'_i\right]^{-1}\left[\sum_{i=1}^{N}W_iD\underset{\sim}{y}_i\right]\right\}, \tag{20}$$

where A is a $(T-1)\times(T-1)$ matrix with 2 on the diagonal elements, -1 on the elements above and below the diagonal elements, and 0 elsewhere.

The GMM estimator has the advantage that it is consistent and asymptotically normally distributed whether α_i is treated as fixed or random because it eliminates α_i from the specification. However, the number of moment conditions increases at the order of T_2, which can create severe downward bias in finite sample (Zilak 1997). An alternative is to use a (quasi-) likelihood approach which has the advantage of having a fixed number of orthogonality conditions independent of the sample size. It also has the advantage of making use of all the available samples, hence can yield a more efficient estimator than (20) (for example, Hsiao et al. 2002; Binder et al. 2005). Since there is no reason to assume the datagenerating process of initial observations, y_{i0}, to be different from the rest of y_{it}, the likelihood approach has to formulate the joint

likelihood function of $(y_{i0}, y_{i1}, \ldots, y_{iT})$ (or the conditional likelihood function $(y_{i1}, \ldots, y_{iT}|y_{i0})$). However, y_{i0} depends on previous values of $\underset{\sim}{x}\ i, -j$ and α_i, which are unavailable. Bhargava and Sargan (1983) suggest circumscribing this missing data problem by conditioning y_{i0} on $\underset{\sim}{x}_i$ and α_i if α_i is treated as random, while Hsiao et al. (2002) propose conditioning $(y_{i1} - y_{i0})$ on the first difference of $\underset{\sim}{x}_i$ if α_i is treated as a fixed constant.

Random Versus Fixed Effects Specification

The advantages of random effects (RE) specifications are as follows:

1. The number of parameters stays constant when sample size increases.
2. It allows the derivation of efficient estimators that make use of both within- and between-(group) variation.
3. It allows the estimation of the impact of time-invariant variables.

The disadvantages of RE specification are that it typically assumes that the individual- and/or time-specific effects are randomly distributed with a common mean and are independent of $\underset{\sim}{x}_{it}$. If the effects are correlated with $\underset{\sim}{x}_{it}$ or if there is a fundamental difference among individual units, that is, conditional on $\underset{\sim}{x}_{it}$, y_{it} cannot be viewed as a random draw from a common distribution, the common RE model is mis-specified and the resulting estimator is biased.

The advantages of fixed effects (FE) specification are that it allows the individual-and/or time-specific effects to be correlated with explanatory variables $\underset{\sim}{x}_{it}$. Neither does it require an investigator to model their correlation patterns.

The disadvantages of the FE specification are as follows:

1. The number of unknown parameters increases with the number of sample observations. In the case when T (or N for λ_t) is finite, it introduces the classical incidental parameter problem (for example, Neyman and Scott 1948).
2. The FE estimator does not allow the estimation of the coefficients that are timeinvariant.

In other words, the advantages of RE specification are the disadvantages of FE specification, and the disadvantages of RE specification are the advantages of FE specification. To choose between the two specifications, Hausman (1978) notes that the FE estimator (or GMM), $\hat{\underset{\sim}{\theta}}_{FE}$, is consistent whether α_i is fixed or random. On the other hand, the commonly used RE estimator (or GLS), $\hat{\underset{\sim}{\theta}}_{RE}$, is consistent and efficient only when α_i is indeed uncorrelated with $\tilde{x}_{it}$. If α_i is correlated with $\underset{\sim}{x}_{it}$, the RE estimator is inconsistent. Therefore, Hausman (1978) suggests using the statistic

$$\left(\hat{\underset{\sim}{\theta}}_{FE} - \hat{\underset{\sim}{\theta}}_{RE}\right)' \left[\operatorname{cov}\left(\hat{\underset{\sim}{\theta}}_{FE}\right) - \operatorname{cov}\left(\hat{\underset{\sim}{\theta}}_{RE}\right)\right]^{-} \left(\hat{\underset{\sim}{\theta}}_{FE} - \hat{\underset{\sim}{\theta}}_{RE}\right) \tag{21}$$

to test RE vs FE specification. The statistic (21) is asymptotically chi-square distributed with degrees of freedom equal to the rank of $\left[\operatorname{cov}\left(\hat{\underset{\sim}{\theta}}_{FE}\right) - \operatorname{cov}\left(\hat{\underset{\sim}{\theta}}_{RE}\right)\right]$.

Nonlinear Models

The introduction of individual-specific effects, α_i, and/or time-specific effects, λ_t, provides a simple way to capture the unobserved heterogeneity across i and over t. However, the likelihood functions are in terms of observables, $(\underset{\sim}{y}_i, \underset{\sim}{x}_i)$, $i = 1, \ldots, N$. Therefore, we will have either to treat α_i as unknown parameters (fixed effects) and consider the conditional likelihood,

$$f\left(\underset{\sim}{y}_i | \underset{\sim}{x}_i, \underset{\sim}{\beta}, \alpha_i\right), i = 1, \ldots, N, \tag{22}$$

or to treat α_i as random and consider the marginal likelihood

$$f\left(\underset{\sim}{y}_i|\,\underset{\sim}{x}_i,\underset{\sim}{\beta}\right)=\int f\left(\underset{\sim}{y}_i|\,\underset{\sim}{x}_i,\underset{\sim}{\beta},\alpha_i\right)f\left(\alpha_i|\,\underset{\sim}{x}_i\right)d\alpha_i, i=1,\ldots,N, \tag{23}$$

where $f(\alpha_i|\underset{\sim}{x}_i)$ denotes the conditional density of α_i given $\underset{\sim}{x}_i$.

When the unobserved individual specific effects, α_i, (and or time-specific effects, λ_t) affect the outcome, *yit*, linearly, one can avoid the consideration of random versus fixed effects specification by eliminating them from the specification through some linear transformation such as the covariance transformation (6) or first difference transformation (18). However, if α_i affects y_{it} nonlinearly, it is not easy to find a transformation that can eliminate α_i. For instance, consider the following binary choice model where the observed y_t takes the value of either 1 or 0 depending on the latent response function

$$y_{it}^{*}=\underset{\sim}{\beta}'\underset{\sim}{x}_{it}+\alpha_i+u_{it}, \tag{24}$$

and

$$y_{it}=\begin{cases}1, \text{if } y_{it}^{*}>0,\\ 0, \text{if } y_{it}^{*}\le 0,\end{cases} \tag{25}$$

where u_{it} is independently, identically distributed with density function $f(u_{it})$. Let

$$y_{it}=E\left(y_{it}|\,\underset{\sim}{x}_{it},\alpha_i\right)+\varepsilon_{it}, \tag{26}$$

then

$$E\left(y_{it}|\,\underset{\sim}{x}_{it},\alpha_i\right)=\int_{-\left(\underset{\sim}{\beta}'\underset{\sim}{x}_{it}+\alpha_i\right)}^{\infty}f(u)du=\left[1-F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i\right)\right]. \tag{27}$$

Since α_i affects $E(y_{it}|\,\underset{\sim}{x}_{it},\alpha_i)$ nonlinearly, α_i remains after taking successive difference of y_{it},

$$y_{it}-y_{i,t-1}=\left[1-F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i\right)\right]-\left[1-F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1}-\alpha_i\right)\right]+\left(\varepsilon_{it}-\varepsilon_{i,t-1}\right). \tag{28}$$

The likelihood function conditional on $\underset{\sim}{x}_i$ and α_i takes the form,

$$\prod_{i=1}^{N}\prod_{t=1}^{T}\left[F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i\right)\right]^{1-y_{it}}\left[1-F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i\right)\right]^{y_{it}}. \tag{29}$$

If T is large, a consistent estimator of $\underset{\sim}{\beta}$ and α_i can be obtained by maximizing (29). If T is finite, there is only limited information about α_i no matter how large N is. The presence of incidental parameters, α_i, violates the regularity conditions for the consistency of the maximum likelihood estimator of $\underset{\sim}{\beta}$.

If $f(\alpha_i|\underset{\sim}{x}_i)$ is known, and is characterized by a fixed dimensional parameter vector, a consistent estimator of $\underset{\sim}{\beta}$ can be obtained by maximizing the marginal likelihood function,

$$\prod_{i=1}^{N}\int\prod_{t=1}^{T}\left[F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i\right)\right]^{1-y_{it}}\left[1-F\left(-\underset{\sim}{\beta}'\underset{\sim}{x}_{it}-\alpha_i\right)\right]^{y_{it}}f\left(\alpha_i|\,\underset{\sim}{x}_i\right)d\alpha_i. \tag{30}$$

However, maximizing (30) involves T-dimensional integration. Butler and Moffitt (1982), Chamberlain (1984), Heckman (1981), and others have suggested methods to simplify the computation.

The advantage of RE specification is that there is no incidental parameter problem. The problem is that $f(\alpha_i|\underset{\sim}{x}_i)$ is in general unknown. If a wrong $f(\alpha_i|\underset{\sim}{x}_i)$ is postulated, maximizing the wrong likelihood function will not yield a consistent estimator of $\underset{\sim}{\beta}$. Moreover, the derivation of marginal likelihood through multiple integration may be

computationally infeasible. The advantage of FE specification is that there is no need to specify $f(\alpha_i|\underset{\sim}{x}_i)$. The likelihood function will be the product of individual likelihood (for example, (29)) if the errors are assumed i.i.d. The disadvantage is that it introduces incidental parameters.

A general approach to estimating a model involving incidental parameters is to find transformations to transform the original model into a model that does not involve incidental parameters. Unfortunately, there is no general rule available for nonlinear models. One has to explore the specific structure of a nonlinear model to find such a transformation. For instance, if $f(u)$ in (24) is logistic, then

$$\text{Prob}\left(y_{it}=1|\underset{\sim}{x}_{it},\alpha_i\right)=\frac{e^{\underset{\sim}{\beta}'\underset{\sim}{x}_{it}+\alpha_i}}{1+e^{\underset{\sim}{\beta}'\underset{\sim}{x}_{it}+\alpha_i}}. \tag{31}$$

Since, in a logit model, the denominators of Prob $(y_{it}=1|\underset{\sim}{x}_{it},\alpha_i)$ and Prob $(y_{it}=0|\underset{\sim}{x}_{it},\alpha_i)$ are identical and the numerator of any sequence $\{y_{i1},\ldots,y_{iT}\}$ with $\sum_{t=1}^{T}y_{it}=s$ always equal to $\exp(\alpha_i s)\cdot\exp\left\{\sum_{t=1}^{T}\left(\underset{\sim}{\beta}'\underset{\sim}{x}_{it}\right)y_{it}\right\}$, the conditional likelihood function conditional on $\sum_{t=1}^{T}y_{it}=s$ will not involve the incidental parameters α_i. For instance, consider the simple case that $T=2$, then

$$\begin{aligned}&\text{Prob}(y_{i1}=1,y_{i2}=0|y_{i1}+y_{i2}=1)\\&=\frac{e^{\underset{\sim}{\beta}'\underset{\sim}{x}_{i1}}}{e^{\underset{\sim}{\beta}'\underset{\sim}{x}_{i1}}+e^{\underset{\sim}{\beta}'\underset{\sim}{x}_{i2}}}=\frac{1}{1+e^{\underset{\sim}{\beta}'\Delta\underset{\sim}{x}_{i2}}}\end{aligned} \tag{32}$$

and

$$\begin{aligned}&\text{Prob}(y_{i1}=0,y_{i2}=1|y_{i1}+y_{i2}=1)\\&=\frac{e^{\underset{\sim}{\beta}'\Delta\underset{\sim}{x}_{i2}}}{1+e^{\underset{\sim}{\beta}'\Delta\underset{\sim}{x}_{i2}}},\end{aligned} \tag{33}$$

(Chamberlain 1980; Hsiao 2003).

This approach works because of the logit structure. In the case when $f(u)$ is unknown, Manski (1987) exploits the latent linear structure of (24) by noting that, for given i,

$$\begin{aligned}&\underset{\sim}{\beta}'\underset{\sim}{x}_{it}>\underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1}\Leftrightarrow E\left(y_{it}|\underset{\sim}{x}_{it},\alpha_i\right)>E\left(y_{i,t-1}|\underset{\sim}{x}_{i,t-1},\alpha_i\right),\\&\underset{\sim}{\beta}'\underset{\sim}{x}_{it}=\underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1}\Leftrightarrow E\left(y_{it}|\underset{\sim}{x}_{it},\alpha_i\right)=E\left(y_{i,t-1}|\underset{\sim}{x}_{i,t-1},\alpha_i\right),\\&\underset{\sim}{\beta}'\underset{\sim}{x}_{it}<\underset{\sim}{\beta}'\underset{\sim}{x}_{i,t-1}\Leftrightarrow E\left(y_{it}|\underset{\sim}{x}_{it},\alpha_i\right)<E\left(y_{i,t-1}|\underset{\sim}{x}_{i,t-1},\alpha_i\right),\end{aligned} \tag{34}$$

and suggests maximizing the objective function

$$H_N(b)=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}sgn\left(\underset{\sim}{b}'\Delta\underset{\sim}{x}_{it}\right)\Delta y_{it}, \tag{35}$$

where $sgn(w)=1$ if $w>0$, $=0$ if $w=0$, and -1 if $w<0$. The advantage of the Manski (1987) maximum score estimator is that it is consistent without the knowledge of $f(u)$. The disadvantage is that (34) holds for any $c\,\underset{\sim}{\beta}$ where $c>0$. Only the relative magnitude of the coefficients can be estimated with some normalization rule, say $||\underset{\sim}{\beta}||=1$. Moreover, the speed of convergence is considerably slower ($N_{1/3}$) and the limiting distribution is quite complicated. Horowitz (1992) and Lee (1999) have proposed modified estimators that improve the speed of convergence and are asymptotically normally distributed.

Other examples of exploiting specific structure of nonlinear models to eliminate the effects of incidental parameters α_i include dynamic discrete choice models (Chamberlain 1993; Honoré and Kyriazidou 2000; Hsiao et al. Hsiao et al. 2005a, b), symmetrically trimmed least squares estimator for truncated and censored data (tobit models) (Honoré 1992), sample selection models (or type II tobit models) (Kyriazidou 1997), and so on. However, often they impose very severe restrictions on the data such that not much of it can be utilized to obtain parameter estimates. Moreover, there are models that do not appear to yield consistent estimator when T is finite.

An alternative to consistent estimators is to consider bias-reduced estimators. The advantage of such an approach is that the bias-reduced estimators may still allow the use of all the sample information so that, from a mean square error point of view, the bias-reduced estimator may still dominate consistent estimators because the latter often have to throw away a lot of the sample, and thus tend to have large variances.

Following the ideas of Cox and Reid (1987), Arellano (2001) and Carro (2006) propose to derive the modified MLE by maximizing the modified log-likelihood function

$$L*\left(\underset{\sim}{\beta}\right)=\sum_{i=1}^{N}\left[\ell_i^*\left(\underset{\sim}{\beta},\hat{\alpha}_i\left(\underset{\sim}{\beta}\right)\right)-\frac{1}{2}\log\ell^*_{i,\alpha_i\alpha_i}\left(\underset{\sim}{\beta},\hat{\alpha}_i\left(\underset{\sim}{\beta}\right)\right)\right] \tag{36}$$

where $\ell_i^*\left(\underset{\sim}{\beta},\hat{\alpha}_i\left(\underset{\sim}{\beta}\right)\right)$ denotes the concentrated log-likelihood function of $\underset{\sim}{y}_i$ after substituting the MLE of α_i in terms of $\underset{\sim}{\beta},\hat{\alpha}_i\left(\underset{\sim}{\beta}\right)$ (that is, the solution of $\frac{\partial \log L}{\partial \alpha_i}=0$ in terms of $\underset{\sim}{\beta}$, $i=1,\ldots,N$) into the log-likelihood function and $\ell^*_{i,\alpha_i,\alpha_i}\left(\underset{\sim}{\beta},\hat{\alpha}_i\left(\underset{\sim}{\beta}\right)\right)$ denotes the second derivative of ℓ_i^* with respect to α_i. The bias correction term is derived by noting that to the order of $(1/T)$ the first derivative of ℓ_i^* with respect to $\underset{\sim}{\beta}$ converges to $\frac{1}{2}\frac{E\left[\ell^*_{i,\beta\alpha_i\alpha_i}\left(\underset{\sim}{\beta},\alpha_i\right)\right]}{E\left[\ell^*_{i,\alpha_i\alpha_i}\left(\underset{\sim}{\beta},\alpha_i\right)\right]}$. By subtracting the order (1/T) bias from the likelihood function, the modified MLE is biased only to the order of $(1/T_2)$, without increasing the asymptotic variance.

Monte Carlo experiments conducted by Carro (2006) have shown that, when $T=8$, the bias of modified MLE for dynamic probit and logit models is negligible. Another advantage of the Arellano–Carro approach is its generality. For instance, a dynamic logit model with time dummy explanatory variable does not meet the Honoré and Kyriazidou (2000) conditions for generating consistent estimators, but will not affect the asymptotic properties of the modified MLE.

Modelling Cross-Sectional Dependence

Most panel studies assume that, apart from the possible presence of individual invariant but period-varying time-specific effects, λ_t, the effects of omitted variables are independently distributed across cross-sectional units. However, often economic theory predicts that agents take actions that lead to interdependence among themselves. For example, the prediction that risk-averse agents will make insurance contracts allowing them to smooth idiosyncratic shocks implies dependence in consumption across individuals. Ignoring cross-sectional dependence can lead to inconsistent estimators, in particular when T is finite (for example, Hsiao and Tahmiscioglu 2005). Unfortunately, contrary to the time series data in which the time label gives a natural ordering and structure, general forms of dependence for cross-sectional dimension are difficult to formulate. Therefore, econometricians have relied on strong parametric assumptions to model cross-sectional dependence. Two approaches have been proposed to model cross-sectional dependence: economic distance (or a spatial approach) and a factor approach.

In regional science, correlation across cross-section units is assumed to follow a certain spatial ordering, that is, dependence among cross-sectional units is related to location and distance, in a geographic or more general economic or social network space (for example, Anselin 1988; Anselin and Griffith 1988; Anselin et al. 2006). A known spatial weights matrix, $W=(w_{ij})$, an $N\times N$ positive matrix in which the rows and columns correspond to the cross-sectional units, is specified to express the prior strength of the interaction between individual (location) i (in the row of the matrix) and individual (location) j (column), w_{ij}. By convention, the diagonal elements, $w_{ii}=0$. The weights are often standardized so that the sum of each row, $\sum_{j=1}^{N} w_{ij}=1$.

The spatial weight matrix, W, is often included into a model specification to the dependent variable, to the explanatory variables, or to the error

term. For instance, a *spatial lag* model for the $NT \times 1$ variable $\underset{\sim}{y} = \left(\underset{\sim}{y}_1', ..., \underset{\sim}{y}_N'\right)', \underset{\sim}{y}_i = (y_{i1}, ..., y_{iT})'$, may take the form

$$y = \rho(W \otimes I_T)\underset{\sim}{y} + X\underset{\sim}{\beta} + \underset{\sim}{u} \tag{37}$$

where X and $\underset{\sim}{u}$ denote the $NT \times 1$ explanatory variables and $NT \times 1$ vector of error terms, respectively, and $\otimes$ denotes the Kronecker product. A *spatial error* model may take the form

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{v} \tag{38}$$

where $\tilde{v}$ may be specified as in a *spatial autoregressive* form,

$$\underset{\sim}{v} = \theta(W \otimes I_T)\underset{\sim}{v} + \underset{\sim}{u}, \tag{39}$$

or a spatial moving average form,

$$\underset{\sim}{v} = \gamma(W \otimes I_T)\underset{\sim}{u} + \underset{\sim}{u}. \tag{40}$$

The spatial model can be estimated by the instrumental variables (GMM estimator) or the maximum likelihood method. However, the approach of defining cross-sectional dependence in terms of 'economic distance' measure requires that the econometricians have information regarding this 'economic distance'. Another approach to model cross-sectional dependence is to assume that the error of a model, say model (39), follows a linear factor model,

$$v_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + u_{it}, \tag{41}$$

where $\underset{\sim}{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of random factors, $\underset{\sim}{b}_i' = (b_{i1}, ..., b_{ir})$, is $r \times 1$ non-random factor loading coefficients, u_{it}, represents the effects of idiosyncratic shocks which is independent of $\underset{\sim}{f}_t$ and is independently distributed across i. (for example, Bai and Ng 2002; Moon and Perron 2004; Pesaran 2006). The conventional time-specific effects model is a special case of (41) when $r = 1$ and $b_i = b_\ell$ for all i and ℓ.

The factor approach requires considerably less prior information than the economic distance approach. Moreover, the number of time-varying factors, r, and factor load matrix $B = (b_{ij})$ can be empirically identified if both N and T are large. However, when T is large, one can estimate the covariance between i and j, σ_{ij}, by $\frac{1}{T}\sum_{t=1}^{T} \hat{v}_{it}\hat{v}_{jt}$ directly, then apply the generalized least squares method, where $\hat{v}_{it}$ is some preliminary estimate of v_{it}.

Large-N and Large-T Panels

Our discussion has been mostly focusing on panels with large N and finite T. There are panel data sets, like the Penn-World tables, covering different individuals, industries and countries over long periods. In general, if an estimator is consistent in the fixed-T, large-N case, it will remain consistent if both N and T tend to infinity. Moreover, even in the case that an estimator is inconsistent for fixed T and large N (say, the MLE of dynamic model (11) or fixed effects probit or logit models (27)), it can become consistent if T also tends to infinity. The probability limit of an estimator, in general, is identical irrespective of how N and T tend to infinity. However, the properly scaled limiting distribution may depend on how the two indexes, N and T, tend to infinity.

There are several approaches for deriving the limits of large-N, large-T panels:

1. *Sequential limits*. First, fix one index, say N, and allow the other, say T, to go to infinity, giving an intermediate limit, then let N go to infinity.
2. *Diagonal-path limits*. Let the two indexes, N and T, pass to infinity along a specific diagonal path, say $T = T(N)$ as $N \to \infty$.
3. *Joint limits*. Let N and T pass to infinity simultaneously without placing specific diagonal path restrictions on the divergence.

In many applications, sequential limits are easy to derive. However, sometimes sequential limits can give misleading asymptotic results. A joint limit will give a more robust result than either a sequential limit or a diagonal-path limit, but will also be substantially more difficult to derive and will apply only under stronger conditions, such as the existence of higher moments. Phillips and Moon (1999) have given a set of sufficient conditions that ensures that sequential limits are equivalent to joint limits.

When T is large, there is a need to consider serial correlations more generally, including both short-memory and persistent components. For instance, if unit roots are present in y and x (that is, both are integrated of order 1) but are not cointegrated, Phillips and Moon (1999) show that, if N is fixed but $T \to \infty$, the least squares regression of y on x is a non-degenerate random variable that is a functional of Brownian motion that does not converge to the long-run average relation between y and x, but it does if N also tends to infinity. In other words, the issue of spurious regression will not arise in a panel with large N (for example, Kao 1999).

Both theoretical and applied researchers have paid a great deal of attention to the unit root and spurious regression properties of variables. When N is finite and T is large, standard time-series techniques can be used to derive the statistical properties of panel data estimators. When N is large and cross-sectional units are independently distributed across i, central limit theorems can be invoked along the cross-sectional dimension. Asymptotically normal estimators and test statistics (with suitably adjustment for finite T bias) for unit roots and cointegration have been proposed (for example, Baltagi and Kao 2000; Im et al. 2003; Levin et al. 2002). They, in general, gain statistical power over their standard time series counterpart (for example, Choi 2001).

When both N and T are large and cross-sectional units are not independent, a factor analytic framework of the form (41) has been proposed to model crosssectional dependency and variants of unit root tests are proposed (for example, Moon and Perron 2004). However, the implementation of those panel unit root tests is quite complicated. When $N \to \infty$, $\frac{1}{N}\sum_{i=1}^{N} u_{it} \to 0$(41) implies that $\overline{v}_t = \underset{\sim}{b}' \underset{\sim}{f}$, where $\underset{\sim}{b}'$ is the cross-sectional average of $\underset{\sim}{b}_i' = (b_{i1},\ldots,b_{ir})$. Approximating $\underset{\sim}{b}_i'\underset{\sim}{f}_t$ by its crosssectional mean function, Pesaran (2005, 2006) suggests a simple approach to filter out the cross-sectional dependency by augmenting the cross-sectional means, $\overline{y}_t$ and $\underset{\sim}{\overline{x}}_t$ to the regression model (38),

$$y_{it} = \underset{\sim}{x}_{it}'\underset{\sim}{\beta} + \alpha_i + \overline{y}_t c_i + \underset{\sim}{\overline{x}}_t' \underset{\sim}{d}_1 + e_{it}, \tag{42}$$

or $\overline{y}_t$, $\Delta\overline{y}_{t-j}$ to the Dickey and Fuller (1979) type regression model,

$$\Delta y_{it} = \alpha_i + \delta_i t + \gamma_i y_{i,t-1} + \sum_{\ell=1}^{Pi} \varphi_{i\ell}\Delta y_{i,t-\ell} + c_i\overline{y}_{t-1} + \sum_{\ell=1}^{Pi} d_{i\ell}\Delta\overline{y}_{t-\ell} + e_{it}, \tag{43}$$

for testing of unit root, where $\overline{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$, $\underset{\sim}{\overline{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \underset{\sim}{x}_{it}$, $\Delta\overline{y}_{t-j} = \frac{1}{N}\sum_{i=1}^{N}\Delta y_{i,t-j}$ and $\Delta = (1-L)$, L denotes the lag operator. The resulting pooled estimator will again be asymptotically normally distributed.

When cross-sectional dependency is of unknown form, Chang (2002) suggests using nonlinear transformations of the lagged level variable, $y_{i,t-1}$, $F(y_{i,t-1})$, as instrumental variables (IV) for the usual augmented Dickey and Fuller (1979) type regression. The test static for the unit root hypothesis is simply defined as a standardized sum of individual IV t-ratios. As long as $F(\cdot)$ is regularly integrable, say $F(y_{i,t-1}) = y_{i,t-1}e^{-c_i|y_{i,t-1}|}$, where c_i is a positive constant, the product of the nonlinear instruments $F(y_{i,t-1})$ and $F(y_{i,t-})$ from different cross-sectional units i and j are asymptotically uncorrelated, even the variables $y_{i,t-1}$ and $y_{i,t-1}$ generating the instruments are correlated. Hence, the usual central limit theorems can be invoked and the standardized sum of individual IV t-ratios is asymptotically normally distributed.

For further review of the literature on unit roots and cointegration in panels, see Breitung and Pesaran (2006) and Choi (2006).

Concluding Remarks

In this paper we have tried to provide a summary of the advantages of using panel data and the fundamental issues of panel data analysis. Assuming that the heterogeneity across cross-sectional units and over time that is not captured by the observed variables can be captured by period-invariant individual specific and/or individual-invariant time-specific effects, we surveyed the fundamental methods for the analysis of linear static and dynamic models. We have also discussed difficulties in analysing nonlinear models and modelling cross-sectional dependence. There are many important issues, such as the modelling of joint dependence or simultaneous equations models, time-varying parameter models (for example, Hsiao 1996, 2003; Hsiao and Pesaran 2006), unbalanced panel, measurement errors (Griliches and Hausman 1986; Wansbeek and Koning 1989), and so on, that were not discussed, but can be found in Arellano (2003), Baltagi (2001) or Hsiao (2003).

Although panel data offer many advantages, they are no panacea. The power of panel data to isolate the effects of specific actions, treatments or more general policies depends critically on the compatibility of the assumptions of statistical tools with the data-generating process. In choosing the proper method for exploiting the richness and unique properties of the panel, it might be helpful to keep the following questions in mind. First, in investigating economic issues what advantages do panel data offer us over data-sets consisting of a single cross section or time series? Second, what are the limitations of panel data and the econometric methods that have been proposed for analysing such data? Third, when using panel data, how can we increase the efficiency of parameter estimates? Fourth, are the assumptions underlying the statistical inference procedures and the data-generating process compatible?

Acknowledgment I would like to thank Steven Durlauf for helpful comments.

Bibliography

Ahn, S.C., and P. Schmidt. 1995. Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68: 5–27.

Anderson, T.W., and C. Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 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.

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

Anselin, L., and D.A. Griffith. 1988. Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association* 65: 11–34.

Anselin, L., J. Le Gallo, and H. Jayet. 2006. 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.

Arellano, M. 2001. Discrete choice with panel data. Working paper no. 0101. Madrid: CEMFI.

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 O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68: 29–51.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70: 91–121.

Balestra, P., and M. Nerlove. 1966. Pooling cross-section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica* 34: 585–612.

Baltagi, B.H. 2001. *Econometric analysis of panel data*. 2nd ed. New York: Wiley.

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

Ben-Porath, Y. 1973. Labor force participation rates and the supply of labor. *Journal of Political Economy* 81: 697–704.

Bhargava, A., and J.D. Sargan. 1983. Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51: 1635–1659.

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.

Breitung, J., and M.H. Pesaran. 2006. Unit roots and cointegration in panels. In *The econometrics of panel data: Fundamentals and recent developments in theory and practice*, 3rd ed., ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Butler, J.S., and R. Moffitt. 1982. A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica* 50: 761–764.

Carro, J.M. 2006. Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics* (forthcoming).

Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* 47: 225–238.

Chamberlain, G. 1984. Panel data. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligato, Vol. 2. Amsterdam: North-Holland.

Chamberlain, G. 1993. *Feedback in panel data models*. Mimeo: Department of Economics, Harvard University.

Chang, Y. 2002. Nonlinear IV unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* 110: 261–292.

Choi, I. 2001. Unit root tests for panel data. *Journal of International Money and Finance* 20: 249–272.

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

Cox, D.R., and N. Reid. 1987. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, B* 49: 1–39.

Dickey, D.A., and W.A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–431.

Granger, C.W.J. 1990. Aggregation of time-series variables: A survey. In *Disaggregation in econometric modeling*, ed. T. Barker and M.H. Pesaran. London: Routledge.

Griliches, Z. 1967. Distributed lags: A survey. *Econometrica* 35: 16–49.

Griliches, Z., and J.A. Hausman. 1986. Errors-in-variables in panel data. *Journal of Econometrics* 31: 93–118.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Heckman, J.J. 1981. Statistical models for discrete panel data. In *Structural analysis of discrete data with econometric applications*, ed. C.F. Manski and D. McFadden. Cambridge: MIT Press.

Honoré, B. 1992. Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60: 533–567.

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

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60: 505–531.

Hsiao, C. 1996. Random coefficient models. In *The econometrics of panel data*, 2nd ed., ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

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

Hsiao, C., and M.H. Pesaran. 2006. Random coefficients 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. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics* 109: 107–150.

Hsiao, C., Y. Shen, and H. Fujiki. 2005b. Aggregate vs disaggregate data analysis – A paradox in the estimation of money demand function of Japan under the low interest rate policy. *Journal of Applied Econometrics* 20: 579–601.

Hsiao, C., Y. Shen, B. Wang, and G. Weeks. 2005a. *Evaluating the effectiveness of Washington State repeated job search services on the employment rate of prime-age female welfare recipients*. Mimeo: University of Southern California.

Hsiao, C., and A.K. Tahmiscioglu. 2005. Estimation of dynamic panel data models with both individual and time specific effects. Mimeo.

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

Kao, C. 1999. Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics* 90: 1–44.

Kyriazidou, E. 1997. Estimation of a panel data sample selection model. *Econometrica* 65: 1335–1364.

Lee, M.J. 1999. A root-N-consistent semiparametric estimator for related effects binary response panel data. *Econometrica* 67: 427–433.

Levin, A., C. Lin, and J. Chu. 2002. Unit root tests in panel data: Asymptotic and finite sample properties. *Journal of Econometrics* 108: 21–24.

Lewbel, A. 1994. Aggregation and simple dynamics. *American Economic Review* 84: 905–918.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life cycle setting. *Journal of Political Economy* 89: 1059–1085.

Manski, C.F. 1987. Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55: 357–362.

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

Neyman, J., and E.L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16: 1–32.

Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49: 1399–1416.

Pakes, A., and Z. Griliches. 1984. Estimating distributed lags in short panels with an application to the

specification of depreciation patterns and capital stock constructs. *Review of Economic Studies* 51: 243–262.
Pesaran, M.H. 2005. A simple panel unit root test in the presence of cross-section dependence. DAE working paper no. 0346, Cambridge University.
Pesaran, M.H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74: 967–1012.
Phillips, P.C. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33: 311–340.
Phillips, P.C., and H.R. Moon. 1999. Linear regression limit theory for nonstationary panel data. *Econometrica* 67: 1057–1111.
Rao, C.R. 1973. *Linear statistical inference and its applications*. 2nd ed. NewYork: Wiley.
Wansbeek, T.J., and R.H. Koning. 1989. Measurement error and panel data. *Statistica Neerlandica* 45: 85–92.
Zilak, J.P. 1997. Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business and Economic Statistics* 15: 419–431.

Lorenz Curve

Nanak Kakwani

JEL Classifications
I3

The Lorenz curve is the most widely used technique to represent and analyse the size distribution of income and wealth. The curve plots cumulative proportion of income units and the cumulative proportion of income received when income units are arranged in ascending order of their income. Max Otto Lorenz, a statistician (born 19 September 1876 in Burlington, USA; retired 1944), proposed this curve in 1905 in order to compare and analyse inequalities of wealth in a country during different epochs, or in different countries during the same epoch – and since then, the curve has been widely used as a convenient graphical device to summarize the information collected about the distributions of income and wealth.

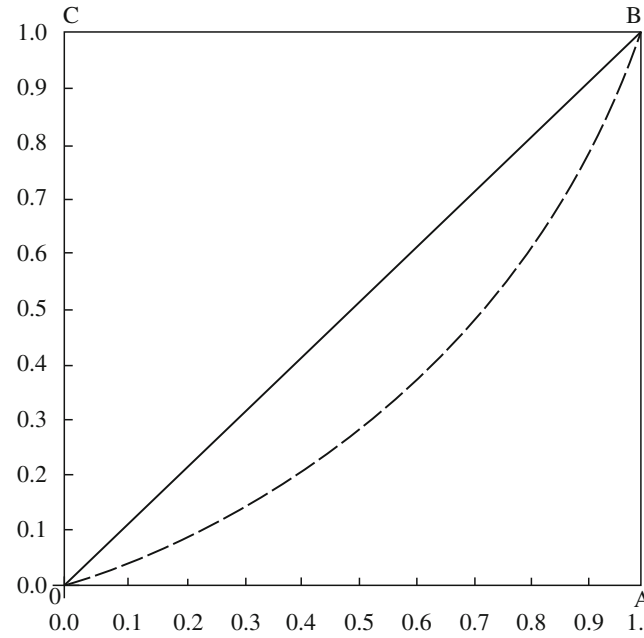
The Lorenz curve may be represented by a function $L(p)$, which is interpreted as the fraction of total income received by the lowest pth fraction of income units. It satisfies the following conditions (Kakwani 1980):

(a) if $p = 0$, $L(p) = 0$
(b) if $p = 1$, $L(p) = 1$
(c) $L'(p) = (x/\mu) \geq 0$ and $L''(p) = (1/\mu\ f(x)) > 0$
(d) $L(p) \geq p$

where income x of a unit (which can be negative for some units but is assumed to be non-negative here for notational convenience) is a random variable with the probability density function $f(x)$ with mean μ and $L'(p)$ and $L''(p)$ are the first and second derivatives of $L(p)$ with respect to p, respectively.

A hypothetical Lorenz curve is illustrated in Fig. 1. The ordinate and abscissa of the curve are $L(p)$ and p, respectively. The slope of the Lorenz curve is positive and increases monotonically, in other words, the curve is convex to the p-axis. From this it follows that $L(p) < p$. The straight line represented by the equation $L(p) = p$, is called the egalitarian line. The curve lies below this line. If, however, the curve coincides with the egalitarian line, it means that each unit receives the same income, which is the case of perfect equality of incomes. In the case of perfect inequality of incomes, the Lorenz curve coincides with OA and AB, which implies that all income is received by only one unit.

Since the Lorenz curve displays the deviation of each individual income from perfect equality, it captures, in a sense, the essence of inequality. The nearer the Lorenz curve is to the egalitarian line, the more equal the distribution of income will be. Consequently, the Lorenz curve could be used as a criterion for ranking income distributions: for if the Lorenz curve for one distribution, X, lies everywhere above that for another distribution, Y, then the distribution X may be said to be more equal than the distribution Y. However, the ranking provided by the curve is only partial – when two Lorenz curves intersect, neither distribution can be said to be more equal than the other. This partial ranking (or quasi-ordering as Sen (1973) calls it)



Lorenz Curve, Fig. 1 The Lorenz curve

need not, however, be considered a weakness of the Lorenz curve. In fact Sen (1973) criticizes the inequality measures that provide complete orderings on the grounds that 'the concept of inequality has different facets which may point in different directions and sometimes a total ranking cannot be expected to emerge'. According to him, the concept of inequality is essentially a question of partial ranking and the Lorenz curve is consistent with such a notion of inequality.

Is there any relation between the Lorenz curve ranking of distributions and social welfare? The answer has been provided by Atkinson (1970) who proved a theorem which shows that if social welfare is the sum of the individual utilities and every individual has an identical utility function which is concave, the ranking of distributions according to the Lorenz curve criterion is identical to the ranking implied by the social welfare function, provided the distributions have the same mean income and their Lorenz curves do not intersect. This theorem implies that one can judge between the distributions without knowing the form of the utility function except that it is increasing and concave. If the Lorenz curves do intersect, however, two utility functions that will rank the distributions differently can always be found.

Atkinson's theorem is based on the assumption that the social welfare function is equal to the sum of individual utilities and that every individual has the same utility function. These assumptions are somewhat limited and have been criticized by DasGupta et al. (1973) as well as by Rothschild and Stiglitz (1973), who have demonstrated that the result is, in fact, more general and would hold for any symmetric welfare function that is quasi-concave.

The Lorenz curve makes distributional judgements independently of the size of income, which as Sen (1973) points out, 'will make sense only if the relative ordering of welfare levels of distributions were strictly neutral to the operation of multiplying everybody's income by a given number'. This is rather an extreme requirement because social welfare depends on both size and the distribution of income.

Working independently on extensions of the Lorenz partial ordering, Shorrocks (1983) and Kakwani (1984) arrived at a criterion which would rank any two distributions with different mean incomes. The new criterion is given by $L(\mu,$

p), which is the product of the mean income μ and the Lorenz curve $L(p)$, whereas the Lorenz curve ranking is based only on $L(p)$. Ranking the distributions according to $L(\mu, p)$ will be identical to the Lorenz ranking if the distributions have the same mean income. This criterion of ranking has been justified from the welfare point of view in terms of several alternative classes of social welfare functions. Kakwani (1984) has used this criterion for international comparison of welfare using data from 72 countries.

As pointed out in the beginning, the Lorenz curve technique was devised as a convenient graphical method to represent and analyse the size distributions of income and wealth. The technique has proved to be extremely powerful and its applications in many areas of applied economics have recently been explored. In analysing data on consumer expenditures Mahalanobis (1960) developed a new technique 'Fractile Graphical Analyses' for comparison of socioeconomic groups at different places or points of time. In this paper, he proposed to extend and generalize the concept of the Lorenz curve to deal with problems of consumer behaviour patterns with respect to different commodities. He suggested that generalized Lorenz curves be called concentration curves, and in fact, used them as a convenient graphical device to describe consumption patterns for different commodities based on data from the National Sample Survey of India.

Kakwani (1977, 1980) provided, however, a more general and rigorous treatment of concentration curves in order to study the relationships among the distributions of different economic variables. He proved theorems which have many applications, particularly in the field of public finance where the effect of taxation and public spending of income distribution is analysed. Other areas in which concentration curves can be applied are inflation as it affects income distribution, estimation of Engel elasticities, disaggregation of total inequality by factor components, and economic growth and income distribution. In a later contribution he used concentration curves to explore how the sense of envy felt by individuals affects the optimal tax structure (Kakwani 1985).

See Also

- ▶ Gini Ratio
- ▶ Pareto Distribution
- ▶ Poverty

Bibliography

Atkinson, A.B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Dasgupta, P., A.K. Sen, and D. Starrett. 1973. Notes on the measurement of inequality. *Journal of Economic Theory* 6: 180–187.

Kakwani, N. 1977. Applications of Lorenz curves in economic analysis. *Econometrica* 45: 719–727.

Kakwani, N. 1980. *Inequality and poverty: Methods of estimation and policy applications*. New York: Oxford University Press.

Kakwani, N. 1984. Welfare ranking of income distributions. *Advances in Econometrics* 191–213.

Kakwani, N. 1985. Applications of concentration curves to optimal negative income taxation. *Journal of Quantitative Economics* 1(1).

Lorenz, M.O. 1905. Methods for measuring concentration of wealth. *Journal of the American Statistical Association* 9: 209–219.

Mahalanobis, P.C. 1960. A method of fractile graphical analysis. *Econometrica* 28: 325–351.

Rothschild, M., and J.E. Stiglitz. 1973. Some further results on the measurement of inequality. *Journal of Economic Theory* 6(2): 188–204.

Sen, A. 1973. *On economic inequality*. Oxford: Clarendon Press.

Shorrocks, A.F. 1983. Ranking income distributions. *Economica* 50: 3–18.

Loria, Achille (1857–1943)

G. de Vivo

Born in Mantua, Loria was Professor of economics at Siena, then Padua, and finally Turin; he died near Turin in 1943. In 1919 he was made a member of the Italian Upper House (where he was one of the few to vote against the Fascist government after their murder of G. Matteotti in 1924). He was well known both in Italy and abroad. A correspondent of the British

Economic Association (later the Royal Economic Society), he contributed several notes (mainly on Italian economics and economists, including obituaries of, among others, Pareto, Barone and Pantaleoni) to the *Economic Journal*, and to Palgrave's *Dictionary* (1894–9). His views on economic theory were rather confused. Böhm-Bawerk wrote: 'Loria's cogitations in the field of theory often impress me as being far more imaginative than they are precise, and to be frequently interlarded with very superficial misinterpretations of other economists' opinion' (1914, p. 479; Schumpeter's judgement was less harsh: 1954, p. 856n.). Indeed, Gramsci entitled a section of his prison notebooks 'Lorianismo', and devoted it to recording ludicrously original conceptions. One of these was Loria's idea that by spreading glue on the body of aircraft one could harvest so many birds as to solve the world food problems and free the workers from their dependence on capitalists for subsistence. Loria tried to present himself as the true originator of the doctrine of historical materialism. As Seligman (1907, p. 136n.) wrote, 'that so many critics in England, France, and Italy should have hailed Loria as the originator of [this] doctrine' is 'a singular testimony to the neglect of Marx's writings outside of Germany' (a similar point is also made by Croce 1896, p. 22). Loria's claim was ridiculed by Engels in the Preface and Supplement to Volume III of Marx's *Capital*, and by Croce (1896).

Selected Works

1880. *La rendita fondiaria e la sua elisione naturale*. Milan: Hoepli.
1886. *La teoria economica della costituzione politica*. Turin: Bocca. English trans. from the 2nd French ed, The economic foundations of society. London: Swan Sonnenschein & Co., 1902.
1890. *Studi sul valore della moneta*. Giornale degli Economisti. A bibliography of his writings was compiled by Einaudi (1932).

Bibliography

Croce, B. 1896. *Le teorie storiche del prof. Loria*. As reprinted in *Materialismo storico ed economia marxistica*. Bari: Laterza, 1968.
Einaudi, L. 1932. Bibliografia di Achille Loria. *La Riforma Sociale* No. 5, Supplement.
Schumpeter, J.A. 1954. *History of economic analysis*. London: Allen & Unwin.
Seligman, E.R.A. 1907. *The economic interpretation of history*, 2nd ed. New York: Columbia University Press.
von Böhm-Bawerk, E. 1914. *Capital and interest. History and critique of interest theories*, 3rd ed. South Holland: Libertarian Press, 1959.

Lösch, August (1906–1945)

Wolfgang F. Stolper

Keywords
Business cycles; Distance; Great depression; Location; Lösch, A.; Partial equilibrium; Population cycles; Regional economics; Secular stagnation; Transfer problem

JEL Classifications
B31

Lösch was born on 15 October 1906 in Oehringen (Württ), though he considered Heidenheim (Brenz) his home. He went to school there, studied in Freiburg with Eucken and in Bonn with Schumpeter and Spiethoff. He was twice a Rockefeller Fellow in the United States, where he did most of the theoretical and empirical work on *Die räumliche Ordnung der Wirtschaft* (1939a), published in the United States as the *Economics of Location* in 1954. His Habilitation (that is, his qualification to teach at a university) on population waves and business cycles was accepted but its unpopular conclusions and his known anti-Nazi views prevented him from getting the *venia legendi*, the

actual permission to teach. He found refuge with the Kiel Institut für Weltwirtschaft, where he became chief of his own research group while at the same time suffering from political interference. He wrote a number of reports for the institute, one of which was published with his conclusions reversed. He kept his personal integrity at great personal cost. He died on 30 May 1945 in Ratzeburg (Holstein) of scarlet fever, which his weakened condition could not tolerate. In 1971, the City of Heidenheim honoured his memory by sponsoring biennial international conferences on location problems, establishing a prize for the best theses in the field and, a few years later, a special honour for older scholars in the field.

Although Lösch's first published paper dealt with the transfer problem, and he continued to be interested in international monetary problems, his only other publications in that field are two discussions of the transfer problem and an extensive fragment in the posthumously published 'Theory of Foreign Exchanges'. The two major subjects of his published work were the relation of population and business cycles and, of course, his highly original *Räumliche Ordnung der Wirtschaft*.

The discussions of population problems anticipate many later developments.

Waves of population increase were neither sufficient nor necessary for the explanation of business cycles. With detailed statistics, some going back to the 17th century, Lösch showed that any relation went from business cycles to population waves, much as recent theory suggests. Though Lösch can claim priority there is no evidence that he actually influenced later developments.

The investigations about a declining and ageing population resulted, however, in quite different conclusions from what was then either politically or academically acceptable. The ageing of the population (the German '*Vergreisung*' has sinister overtones absent from the English equivalent) had its economic compensations. It allowed the better training of the younger generation and increased capital accumulation and productivity. Even in military terms, fewer but better trained and better equipped people were preferable to more but less skilled individuals. In short, fewer young people allowed greater savings and investments leading to increased productivity and growth. This differed substantially from the then prevalent secular stagnation thesis and is much more in keeping with the warnings of present-day development economists of the dangers of rapid population growth. Lösch's earlier *Was ist vom Geburtenrückgang zu halten*? (1932) was later put on the index by the Nazis and his doctoral thesis on the same topic was effectively suppressed.

Lösch's greatest contribution dealt, in most general terms, with general equilibrium theory applied to space. Distance itself becomes the central phenomenon. Lösch's intellectual predecessors dealt with this problem essentially in two ways.

They either solved a partial equilibrium system (Alfred Weber) or they substituted a series of smaller regions for one large one (Ohlin).

Going from partial to general equilibrium, and investigating the structure of the region instead of taking it as given, involved the substitution of a very general set of assumptions for the usual *ceteris paribus* assumptions made. In Weber (and practically everyone else) the locations of markets, raw materials and populations are assumed. In Lösch the basic assumption is a perfectly even distribution of population and of all raw materials. With these extraordinarily general and brilliantly unrealistic assumptions Lösch succeeds in showing that competitive forces alone will establish a system of locations which, in turn, can be understood either as agglomerations of productions or the intersection of fewer or more crossroads, all being simultaneously determined.

Lösch presents a Walrasian model with distance built in as a system of coordinates of location. His most famous contribution, however, is the analysis of the *structure* of an economic landscape on the basis of the simple generalized assumptions mentioned. The empirical work related mostly to the American Mid-West, where the assumptions are approximately realistic. One

test of the genius of the model is that, unlike with most theoretical models, the introduction of more realistic assumptions simplifies rather than complicates the model.

In the 'ideal' Lösch landscape the basic unit is a hexagon. This follows from the condition that consumers are initially equidistant from each other, that each producer and consumer must lie within the market area of each good and that there must be no empty corners. Modifications introduced are rectangular areas on the model of, say, the layout of American counties; or the effect of different resource endowments of different areas; or of a border separating what might otherwise be one market area.

The work does not exhaust itself with equilibrium analysis or the structure of economic landscapes. There is a dynamic analytical and empirical study of how business cycles spread over the economic landscape or how transfers are made over and between areas through intra-regional adjustments in connected areas and from one sub-market to another. Thus the initial impact of a change in demand in one landscape capital might first be felt in the capital in the centre of another landscape and spread from there in declining ripples to the border. There is a study of how the Great Depression spread in time and geographically through an area. The usual multiplier is supplemented by a spatial one.

The Lösch analyses the *Gestalt* of a region rather than defining it by such criteria as the immobility of factors of production between but not within regions: all factors are mobile at a cost which varies with distance, even land whose physical immobility is substituted for by changes in its utilization. The case of completely specific resources is investigated, though considered rare.

Lösch left a number of unfinished studies, and plans for many more. His is probably the most original book published on economics in the German language between the two world wars. Most scholars would consider themselves lucky if they had added a layer of bricks to an existing wall. Only few scholars can claim to have started a new wall, and even fewer to have started a new building. Lösch is one of those few scholars.

See Also

▶ Location Theory

Selected Works

1930. Eine Auseinandersetzung über das transfer problem. *Schmollers Jahrbuch* 54: 1193–1206.

1932. *Was ist vom Geburtenrückgang zu halten?* 2 vols. Heidenheim: Privately Published.

1936a. *Bevölkerungswellen und Wechsellagen*. Jena: Gustav Fischer.

1936b. Die Vergreisung wirtschaftlich gesehen. *Schmollers Jahrbuch* 60: 577–685.

1936–7. Population cycles as a cause of business cycles. *Quarterly Journal of Economics* 51: 649–662.

1938. The nature of economic regions. *Southern Economic Journal* 5(1): 71–78.

1939a. *Die räumliche Ordnung der Wirtschaft. Eine Untersuchung über Standort, Wirtschaftsgebiete und Internationalen Handel*. 2nd Rev edn, 1944. 3rd edn (reprint of the 2nd edn), Jena: Gustav Fischer, 1962. 2nd edn trans. as *The economics of location*. New Haven: Yale University Press, 1954.

1939b. Eine neue Theorie des Internationalen Handel. *Weltwirtschaftliches Archiv* 50: 308–328. Trans. as A new theory of international trade. *International Economic Papers* No. 6. London: Macmillan, 1956.

1949. Theorie der Währung. Ein Fragment. *Weltwirtschaftliches Archiv* 62: 35–88.

Bibliography

Riegger, R. ed. 1971. *August Lösch in Memoriam*. Heidenheim: Verlag der Buchhandlung Meuer. Contains eight contributions and a bibliography of 78 items, including literature about Lösch.

Valavanis, S. 1955. Lösch on location. *American Economic Review* 45: 637–644.

Zottmann, A. 1949. Dr. Habil. August Lösch, gestorben am 30. Mai 1945. *Weltwirtschaftliches Archiv* 62(1), 28–31. Appended bibliography, 32–4.

Lotka, Alfred James (1880–1949)

Joel E. Cohen

Abstract

Alfred James Lotka was a many-sided scientist who pioneered the mathematical theory of population. He created the demographic theory of stable populations as a special case of a general theory of renewal. He developed and applied to contemporary demographic data the important concepts of net rate of reproduction and intrinsic rate of natural increase. He created and analysed mathematical models of predation and competition that remain influential in theoretical ecology and proposed a comprehensive physico-chemical view of evolution. He also evaluated the expected value of lifetime earnings of workers of specified ages, analysed mathematical models for the epidemiology of malaria, and calculated rank-size distributions of scientific productivity.

Alfred James Lotka was a many-sided scientist who pioneered the mathematical theory of population. He created the demographic theory of stable populations as a special case of a general theory of renewal. He developed and applied to contemporary demographic data the important concepts of net rate of reproduction and intrinsic rate of natural increase. He created and analysed mathematical models of predation and competition that remain influential in theoretical ecology and proposed a comprehensive physico-chemical view of evolution. He also evaluated the expected value of lifetime earnings of workers of specified ages, analysed mathematical models for the epidemiology of malaria, and calculated rank-size distributions of scientific productivity.

Lotka was born on 2 March 1880 to French-speaking parents of American citizenship in Lemberg, Austria (now Lwów, Ukrainian SSR). He received his early education in France, Germany and England. From Mason College, Birmingham University, he received the BSc in 1901 and the DSc in 1912. During a year's study of chemistry at Leipzig University in 1901–2, he developed concepts for a mathematical theory of evolution.

He came to the United States in 1902 and worked as an industrial chemist. In 1908, he registered at Cornell University as a doctoral candidate in physics and mathematics, but left in 1909 with the degree of AM. After working as an examiner at the US Patent Office, as a physicist at the US Bureau of Standards, as a freelance writer, as an editor of the *Scientific American* Supplement, and as a chemist at the General Chemical Company, he accepted in 1922 a temporary research appointment in Raymond Pearl's Human Biology group at Johns Hopkins University. Between 1922 and 1924, he completed his magnum opus, *Elements of Physical Biology* (1925a), putting flesh on the bones of ideas he had developed over the preceding quarter century.

From 1924 until his retirement in 1948, Lotka worked for the Metropolitan Life Insurance Company in New York City as supervisor of mathematical research in the Statistical Bureau (1924–33), as general supervisor (1933–4), and as assistant statistician (1934–48). He married Romola Beattie on 5 January 1935; they had no children. He was president of the Population Association of America in 1938–9, president of the American Statistical Association in 1943, and vice-president of the International Union for the Scientific Investigation of Population Problems. In retirement, Lotka revised and translated portions of his *Théorie analytique des associations biologiques* (1934, 1939c). After an illness of a few weeks, he died in Red Bank, New Jersey, on 5 December 1949.

Lotka's more than one hundred scientific papers and five books range widely, from the mathematics, physics and chemistry of his early training, to fields some of which he helped to create, including theoretical and applied demography, ecology, epidemiology, other mathematical social sciences including economics, and operations research. He was a gifted and engaging expositor.

In 1907, Lotka analysed homogeneous density-independent populations closed to migration and

growing at a given rate, in which individuals are subject to a given schedule of mortality. Lotka supposed that the age structure, that is, the proportions of individuals in each age group, is independent of time, and expressed the age structure in terms of the given growth rate and mortality schedule. Estimating the growth rate and the mortality schedule from reports of the Registrar General of England and Wales for 1871–80, Lotka showed that the predicted per capita rates of birth and death and predicted age structure agree well with the corresponding observations.

Lotka entitled his major paper of 1907 containing this analysis 'Studies on the mode of growth of material aggregates'. He followed the analysis immediately by a model of isothermal monomolecular reactions. To explain this juxtaposition, he 'recognized the problem of chemical dynamics as a special case of a wider problem: . . . [namely,] the study of the laws governing the distribution of matter among complexes of any specified kind, as determined by their general physical character'. This wider problem, he wrote, 'may be taken to represent the quantitative formulation of the problem of evolution in its most general terms'. He devoted much of the rest of his scientific career to amplifying this perspective of evolution both in abstract generality and in particular contexts.

In 1911, Francis Robert Sharpe (1870–1948) and Lotka showed that the time-invariant age structure, which Lotka had previously analysed, is stable (*see* ▶ Stable Population Theory). These articles of 1907 and 1911 form the core of Lotka's contributions to population analysis.

In 1934 and 1939, Lotka published the two parts of *Théorie analytique des associations biologiques, I: Principes; II: Analyse démographique avec application particulière à l'espèce humaine*. The latter (1939c) summarizes his contributions to the population analysis of a single, principally the human, species. The book treats the theory of stable populations, including the solution of the renewal equation, and of logistically growing populations; the progeny of a population element; indices of population growth; fecundity by birth order and family size; orphanhood and family composition; and the extinction of a line of descent. Nearly half a century later, the central ideas of demography, as they are used by most demographers, remain close to those Lotka codified in 1939.

Starting in 1910, Lotka described chemical reactions that display damped oscillations. In 1920, he discovered a pair of nonlinear first order differential equations that display undamped oscillations and interpreted them both chemically and in terms of plant-herbivore interactions. Independently, in 1926, the Italian mathematician Vito Volterra (1860–1940) published a model for one species feeding on another that is mathematically identical to Lotka's 1920 model of undamped oscillations. Others of Volterra's models fell within the framework of Lotka's general theories for the interactions of chemical or biological species.

Virtually all ecologists now refer to the system $dx/dt = x(abx - cy)$, $dy/dt = y(f - gx - hy)$, where x and y are the abundances or biomasses of two species competing for a common resource, and a, b, c, f, g, h are positive parameters, as the Lotka–Volterra equations. Volterra analysed these equations in 1926 and Lotka in 1932 by different methods.

Lotka concluded a 1932 analysis of the Lotka–Volterra equations by remarking that 'It is perhaps hardly to be expected that concrete examples of the law of growth for two populations here discussed shall be found in nature', a warning many subsequent ecologists who have used the equations have forgotten. He suggested that an experimental realization of the equations would be 'interesting'. At the same time, 1932, the Russian experimentalist G.F. Gause (b. 1910) published microbiological experiments consistent with the Lotka–Volterra equations. The theory of the stability of the Lotka–Volterra equilibria and his own experiments led Gause to formulate, and to attribute to Lotka and Volterra, a principle of competitive exclusion that now bears Gause's name.

Lotka followed contemporary economic thought (he was a member of the Royal Economic Society) and sought to contribute to it. In the same 1932 paper, Lotka suggested that the treatment which has here been developed in the analysis of the growth of multiple populations, may find more immediate application in the field of economics

. . . Cournot's treatment of the problem of competition has been criticized on the ground that . . . any one competitor who should possess the slightest advantage over the others, would ultimately displace them entirely, and hold the field in absolute monopoly.

Lotka showed how the spatial dispersion of competitors avoided this criticism.

Of all his works, Lotka was proudest of *Elements of Physical Biology* (1925a). When the book was reissued in 1956, Herbert A. Simon saw in its exposition of systems of differential equations, stability of equilibria and comparative statics 'many of the sources of Samuelson's analysis of the relations of statics and dynamics in his *Foundations* – a debt which Samuelson acknowledges' (Simon 1959). He added, 'It is easy to show that much that has happened in mathematical social science in the thirty years since the publication of the first edition of *Elements of Physical Biology* lies in directions along which the book points.' His final assessment is one that many readers of Lotka share: 'When I weary of reading Lotka for his mathematics, I read him for sheer delight, and for the broad perspective he gives me of the world in which I live.'

See Also

- ▶ Predator–Prey Models
- ▶ Stability
- ▶ Stable Population Theory
- ▶ Volterra, Vito (1860–1940)

Selected Works

1907. Studies on the mode of growth of material aggregates. *American Journal of Science* 24: 199–216, 375.

1911. (With F.R. Sharpe.) A problem in age-distribution. *Philosophical Magazine* 21: 435–438.

1925a. *Elements of physical biology*. Baltimore: Williams/Wilkins. Reprinted as *Elements of mathematical biology*. New York: Dover, 1956; Lotka's own list of his scientific papers is appended to this reprint, but it is incomplete.

1925b. (With L.I. Dublin.) On the true rate of natural increase as exemplified by the population of the United States, 1920. *Journal of the American Statistical Association* 20: 305–339.

1928. The progeny of a population element. *American Journal of Hygiene* 8(6): 875–901.

1930. (With L.I. Dublin.) *The money value of a man*. New York: Ronald Press Co. Revised edn., 1946.

1934. *Théorie analytique des associations biologiques. Pt I: Principes*. Actualités Scientifiques et Industrielles No. 187. Paris: Hermann et Cie.

1936. (With L.I. Dublin.) *Length of life, a study of the life table*. New York: Ronald Press. Revised edn., 1949 (with L.I. Dublin and M. Spiegelman).

1937. (With L.I. Dublin.) *Twenty-five years of health progress: A study of the mortality experience among the industrial policyholders of the metropolitan life insurance company, 1911 to 1935*. Metropolitan Life Insurance Company. Supplement: Health Progress 1936–1945.

1939a. A contribution to the theory of self-renewing aggregates, with special reference to industrial replacement. *Annals of Mathematical Statistics* 10(1): 1–25.

1939b. On an integral equation in population analysis. *Annals of Mathematical Statistics* 10(2): 144–161.

1939c. *Théorie analytique des associations biologiques. Pt II: Analyse démographique avec application particulière à l'espèce humaine*. Actualités Scientifiques et Industrielles No. 780. Paris: Hermann et Cie.

1948. Application of recurrent series in renewal theory. *Annals of Mathematical Statistics* 19(2): 190–206.

Bibliography

Gause, G.F. 1964. *The struggle for existence*. New York: Hafner.

Kingsland, S.E. 1985. *Modeling nature: Episodes in the history of population ecology*. Chicago: University of Chicago Press.

Notestein, F.W. 1982. Demography in the United States: A partial account of the development of the field. *Population and Development Review* 8(4): 651–687.
Simon, H.A. 1959. Review of *elements of mathematical biology* by Alfred J. Lotka. *Econometrica* 27(3): 493–495.

Low Pay

Frank Wilkinson

In neoclassical theory the market operates in such a way as to equate wages with the marginal product of labour. The productivity of workers is determined largely by their skill and capacity for work and as these vary between individuals so will earnings. However the workings of the market will tend to equalize *efficiency-earnings*, i.e. earnings 'measured . . . with reference to the exertion of ability and efficiency required of the workers' (Marshall 1952, p. 456) and consequently the wages of individuals will be proportionate to their productivity. Low pay is therefore explained by the 'quality' of individuals; a view expressed at its starkest by Hicks: 'Casual labour is often badly paid, not because it gets less than it is worth, but because it is worth so appallingly little' (Hicks 1963, p. 82).

Labour force quality also plays an important part in many explanations of the current high level of unemployment. The concentration of unemployment amongst the old, the young and the low paid is taken as evidence of a supply side constraint which can only be removed by education and training to upgrade workers to fit the jobs available, or by the creation of jobs with a low skill content (Layard 1986).

The emphasis given to labour 'quality' (or more precisely non-quality) in determining low pay is underpinned in the neoclassical scheme by theorizing which equates 'quality' with investment in human capital. Despite the central place it occupies in the neoclassical supply side theory of the labour market, the human capital approach receives little empirical support as a general theory of the structuring of wages or employment opportunities. Years of formal education may help explain why white males reach top positions, but the returns to education and training are much less for women and racial minorities and within large segments of the labour market – manual employment for example – the benefits of education and training may be very low. (Wilkinson 1981, especially Ryan, Rosenberg and Buchele). A second major criticism of the human capital approach is the central focus it gives to *formal* education and training. This leaves completely out of account such socially acquired skills as domestic organization and caring, which are central requirements of many feminized occupations, and use of basic tools and machinery, which form the basis of many manual occupations. Jobs requiring such abilities are consequently labelled 'unskilled'.

But perhaps the most telling criticism of the human capital approach is that it implicitly assumes a matching between education and training, on the one hand, and the content of jobs, on the other. It only takes a moment's reflection to realise how small is the vocational element in most educational courses above the provision of basic literacy and numeracy. However, the idea of substantial education and training as a necessary pre-requisite for the acquisition for a 'good' job is partly salvaged by signalling theory (Spence 1973). Here, in the absence of any more concrete tests, education as well as other social indicators, including sex and race, signals to the employer the suitability of particular applicants for jobs. It is of course true that employers would only retain their faith in signals if their expectations of adequate job performance by the signaller are fulfilled. But because potential employers have no direct evidence on the performance of the job applicant they have no way of judging the capabilities to fulfil the requirements of the job of those applicants *eliminated* by the signalling procedure.

The idea that individuals signal their suitability for certain types of job by a wide range of social indicators suggests that 'quality' is a social category. Empirical research lends weight to this by demonstrating that many jobs in the low pay, low skill category require significant degrees of skill and carry much responsibility (Craig et al. 1985;

Blackburn and Mann 1979); much more so than many jobs in the primary sector. This can also be linked to the demonstration that skill itself is much more a social than a technical classification and largely determined by the power relations between organized labour and employers and between different segments of the labour force (Turner 1962; Rubery 1978). This more empirically based research has tended to show that the structure of job opportunities is dependent on such factors as product market organization, technology, industrial organization and the degree of managerial control of the labour process. Access to the resulting 'good' and 'bad' job is largely determined by social and political organization in which worker ability – either actual or potential – is not the determining factor.

Low Pay and Social Disadvantage

The Point of Departure

As I understand it, orthodox labour market theory rests on three assertions: labour is a scarce resource, individuals are inherently unequal, and they are free to compete for a wide range of jobs. In such circumstances, the market operates to allocate 'scarce means to alternate uses' and provides equality of opportunities; consequently, wage differentials measure the inequality of individuals in terms of the quantity and quality of their labour. It is my contention that none of these assertions is tenable. Labour is in more or less abundant supply, its usage is demand constrained and in terms of the requirements of the vast majority of jobs workers are intrinsically equal. In these circumstances the institutions on both the supply and demand side of the labour market operate in precisely the opposite way to that postulated by orthodox economic theory; they discriminate between equal claimants in the allocation of scarce good jobs and in the process generate wage differentials.

The Structuring of Labour Markets

In a capitalist system labour is inherently weak when compared with capital. This power imbalance may be somewhat redressed if workers have access to resources from domestic or other out-of-market sources or from the state. But generally such resources are not sufficient to allow workers to maintain a reasonable standard of life independent of the labour market. Therefore collective actions are more important than out-of-market resources in redressing the imbalance of power between capital and labour and these are organized in the domestic sector, in the market, and at the level of the state. However, the ability to counter the inherent superiority of capital varies between groups of workers by degrees determined by their access to out-market resources, education and training, by their role in domestic production and by the organization of other groups of workers aimed at their exclusion. The consequent structuring of the potential labour supply at any one level – domestic, market and the state – may be offset or buttressed by organization at another.

Organization at the level of the family and the community allows the withdrawal of certain classes of workers from the labour market – particularly women and children – and provides alternative sources of subsistence and mutual support, which strengthens the bargaining power of individual workers. Moreover the family provides resources for education and training which enhances the market value of individuals. Domestic and community organization are therefore sources of strength but at the same time they serve to sectionalize and fragment the workforce. Patriarchy, which forms the basis for the organization of the family, inherently weakens the position of women in domestic production. The unequal distribution of wealth between families and communities differentiates the labour force in terms of education and training. Moreover, out-of-market and in-market disadvantages are reinforcing. Women's domestic responsibilities, which inhibit labour market participation, and male dominance in the labour market, which ensures an inferior status, place women in a lower paid and easily exploitable category which is exacerbated by their partial dependence on family income for subsistence. The lack of out-of-market resources for education and training discriminates against the young from poor families and communities and the

consequent low levels of pay reinforce their disadvantage.

Worker organization in the labour market is also typified by the contradictory tendency towards collective action and sectionalism. Trade unions (here defined to include professional associations and the more informal 'old boy' networks which create privileged access to, and establish control of, classes of jobs) necessarily operate on a dual principle of representing the common interest of those within the union whilst protecting their areas of influence by policies of exclusion. This need is reinforced by the stratification of the labour force from the supply side which creates a pool of cheap labour posing a continuous threat to organized labour which further encourages demarcation and exclusion strategies.

The state provides the third main force structuring the labour supply. Much of the state's activity can be interpreted as being in the interest of capital in ensuring the existence of a disciplined, trained and healthy labour force and in maintaining a reserve army of labour in readiness for mobilization by the provision of minimal social welfare. On the other hand the struggle between labour and capital at the level of the state has resulted in important gains for labour.

To varying degrees between countries the state has enacted legislation laying down minimum conditions for the employment of labour, removing restrictions on trade union activities and establishing legally binding minimum wages. The state has also intervened by extending state provision into health, education, housing and social security. State pensions, unemployment pay and sickness benefit supplemented and in many cases replaced those provided by trade unions and private insurance and no doubt more informal support from the family and community. But the state system tends to be more comprehensive and more efficient than the combination of provision by the market, charities and intra-family and intra-community transfers, and in particular extended provisions to those partially or totally excluded from private provision.

An important effect of the development of the floor of rights by the state has been to benefit differentially workers in the lowest paid segments of the labour force. The lifting of legal constraints on collective industrial action and minimum wage legislation stood to benefit most those who individually were in the weakest bargaining position. The lowest paid who were least able to provide for education, health and social welfare in the market stood to gain most from universal provision by the State. Thus one of the effects of extension of the welfare state has been to counteract effects of labour market segmentation. Social security has lifted the burden of poverty somewhat; education and improved health care have raised expectations.

However the extent to which the state counteracts segmentation in the labour market should not be exaggerated. The continued existence alongside the state system of private education with access based on the ability to pay but which is nevertheless heavily subsidized by the state ensures privileged access to enhanced job and earning opportunities. Such privilege is not confined to the ability to pay. The middle classes' greater knowledge of the system, their social training and articulateness, a shared social background with officials, and an awareness of the benefits of education and the health service places them in a stronger position than the working class to exploit the state system.

The modern welfare state has also preserved the 19th-century distinction between the 'deserving' and 'undeserving' poor and the notion of 'less eligibility' for those in receipt of state financial support. Social welfare benefits are kept at a minimum so as to maintain the 'incentive' to work and a clear distinction is drawn between benefits secured by contribution and others which are given at the discretion of the social security administration. This discriminates against groups with low pay and uncertain employment, and particularly women who find difficulty in maintaining continuity of employment necessary to secure benefits as of right. In many countries women are further disadvantaged by the fact that the state embodies in the rules for social provision the notion that they are economically dependent on men.

The Operation of Labour Markets

The labour force is therefore stratified by class, race, nationality, religions, sex and many other

factors. These divisions are created and reinforced by discrimination, differential access to education and training, professional associations, trade unions, employers' associations and ratified by social beliefs and conventions. Supply-side structuring has its demand-side counterpart in the hiring rules adopted by firms which rest on signals transmitted by social characteristics (age, sex, race, educational qualification etc.) which are only partially objectively based but which are taken to measure the relative worth of job applicants. Thus the filtering process rations out the scarce good jobs and ensures that the workers with the least attractive features in terms of social classification are employed in the lowest paid and most insecure occupations. The important features of structured labour markets are that relative wages are no guide to relative skills or productivity and that workers of equal skill or potential ability are employed at widely different wage levels.

The general characteristic of labour markets is that workers are not free to move from one job to another. Whilst individuals are free to vacate an existing job, their access to others is severely curtailed. Access to vacant jobs is carefully controlled, and the higher the pay the more restrictive the rules of entry. Rules of exclusion operate on all groups at all levels and are mutually reinforcing in the sense that workers in each labour market group, excluded from better jobs, more carefully protect those within their control. Jobs which are accessible to almost anyone are generally those which almost nobody would want.

Access to particularly jobs, and the incomes associated with them, depends largely on social circumstances as much as ability and qualifications. The social position of married women, for instance, has made them willing to accept jobs which attract relatively low wages and offer poor working conditions. The choice of position in the labour market hierarchy is restricted by social constraints even though on purely economic criteria their productivity would open up a wider selection of occupations. In a similar way access to the small number of 'good' jobs towards the top of the labour market hierarchy is restricted by institutional rules and restrictions which are supported by custom and social acceptance.

The labour market opportunities of individuals depend then, on the one hand, on qualifications, aspirations and information which are determined by upbringing and by education and, on the other hand, the occupational structure which is determined by the interplay of technical and social factors and which determines the level and range of skills and of earning opportunities. Thus what constitutes the labour market varies for individuals. University graduates, for example, will have qualifications which will admit them to the highest level jobs but which will not necessarily exclude them from occupations lower in the hierarchy. Moreover they will be able to adopt a national or even international perspective on job opportunities, whereas the actual or perceived job opportunities of a worker with minimum educational attainments will be confined to a small occupational and geographical territory. The localization of job opportunities will be reinforced by the network of information and contacts by which jobs are secured and by the acquisition of specific skills, seniority rights and job experience which together determine the level of earnings and job security. Such factors serve to trap workers into declining areas and industries to a degree which is inversely related to their ability to retrain and to gain access to jobs which offer prospects comparable to those relinquished (Stedman Jones 1984). Such potential mobility will depend on the individual's ability to signal that he can adjust to changed circumstances and these indicators – for example, educational attainment, age, sex and race – may be determined by social norms and values, unrelated to actual ability and performance.

The Dynamics of the Labour Market

The fact that labour markets are divided into largely non-competing occupational groups does not mean that they are inflexible. There is no historical evidence that the supply of labour has proved to be a long-term constraint on economic growth and development although problems of integrating newly mobilized reserves of labour may have placed a ceiling on the pace of

expansion. Nor is there any lack of evidence of capital's ability to restructure the demand for labour so as to economize in its use, however its ability in this respect may be constrained by labour organization. But the process of labour market restructuring has not been continuous or in a single direction.

During prolonged periods of expansion of demand considerable pressure develops on the stock of labour power with a growing need to expand it in both qualitative and quantitative terms. The response to this comes in the form of an increase in the fraction of the population seeking work – recently in the form of increased employment of arrived women – by an increase in the hours of overtime and multiple job holding, by inter-industry, inter-regional and international shifts and by the 'up-grading' of labour by education and training. Amongst those already in the labour force, flexibility is achieved by the lower tiers serving as reserves of labour for higher tiers and this upgrading is relatively easily achieved because workers are generally underemployed and all that is normally required is a change in hiring rules rather than any radical retraining programme.

In such periods the state will come under increasing pressure to increase expenditure on education and training to facilitate the upgrading of the existing labour force and to induce entry of potential new recruits from outside. Increased employment of workers from the periphery will require increased expenditure to assist recruitment, training and possibly to subsidise transport and accommodation for new recruits. In the case of married women, who provide the most easily mobilized reserve of labour, it may also be necessary to extend child care facilities.

Periods of rapid growth will also generally be periods of rapid change in techniques and industry's structure and location. This will increase the already heavy pressure on government to increase the size and upgrade the labour force by more expenditure on education and training and inducements for a rapid transfer of labour from declining to expanding sectors. Health care and health and safety at work will also be given priority in periods of high and growing employment to maintain the labour force, to ensure the quick return to employment of sick workers, and to extend the working life. In this latter respect in the post-war periods, in conditions of chronic labour shortage, significant tax incentives were given to the old to encourage the postponement of retirement.

One effect of a high level of demand for labour was therefore to induce an increase in state expenditure to improve the labour supply in quantitative and qualitative terms. More generally, high levels of employment have increased political pressure from the trade unions and other pressure groups concerned with poverty for a general improvement in social welfare benefits. Similar tendencies are observable in the development of trade unions. High employment strengthens the bargaining power of trade unions and in particular tends to extend its coverage to incorporate hitherto unorganized workers particularly amongst the lowest paid. This has added to the pressure on government to outlaw discrimination against racial minorities, legislate in favour of equal pay for women and improved employment conditions such as paid maternity leave and the right to reinstatement to jobs after maternity leave. Thus economic, political and social pressures combined in the upgrading of the labour force in such a way as to benefit particularly those at lower levels in the hierarchy.

In periods of high and rising unemployment the upgrading process described above is reversed. Changes in hiring rules dispel the less well qualified from the upper levels of the employment hierarchy and these people in turn shunt new arrivals out of the lower levels. Thus there is a general downgrading of the labour with the burden of increased job uncertainty and unemployment falling on the lowest paid and particularly such disadvantaged groups as the young, the old, women and racial minorities.

Governmental response to the growing crisis of unemployment and underemployment has been to identify as its primary cause egalitarian welfare state expenditure and inflexibility in the labour market. The consequence has been the reversal of policies adopted in the previous upswing. There has been a general reduction in government

expenditure and to 'improve' the working of the labour market out-of-work social benefits have been reduced and access to them made more difficult to discourage the 'work-shy', legislation has been enacted to weaken trade unions and to remove legal minimum wage protection and the obligations on employers to provide job security and health and safety at work have been lifted.

The impact of these changes have fallen disproportionately on the lower tiers of the labour market. In sectors protected from the economic recession and from governments by the effectiveness of social and industrial organization, real income continued to grow and employment has remained secure. However this sector is shrinking in size and individuals are reluctant to change jobs because of the obvious risks involved so that job opportunities are rare. The lower tiers of the labour market have been swollen by the collapse of employment in certain sectors – particularly manufacturing – and the downgrading process outlined above. It is in the lower segment of the labour market that the main weight of economic and social crisis has fallen. It is here that the long-term unemployed are located and where frequent bouts of unemployment are the common experience of those retaining some measure of labour market attachment. Jobs have also become increasingly short term, work more casualized, and wages have fallen substantially in relative terms and frequently in real terms.

The further governmental response to this policy induced sharpening of the division within the labour market, increased unemployment and under-employment and impoverishment is the classic one: to blame the victim for the disease. Social disorder in areas of particular deprivation 'mainly resulting from unemployment are identified as a break down of law-and-order and met by intensified policing. Unemployment is increasingly identified with an unwillingness to work leading to the growing cohesion of the out-of-work backed by the sanction of withdrawal of social benefits and hence total destitution. The claim that the workforce is becoming 'unemployable' has led to a range of make-work which is normally the province of the low paid. Heavily subsidized youth training – designed to raise the 'quality' of labour – are abused as sources of cheap labour and as screening devices which cost little and frequently replace superior forms of industrial training. These measures at best disguise some of the unemployment and at worst serve to fragment and further deprive the most disadvantaged in society.

Conclusions

Low pay results from a shortage of good job opportunities and an unequal distribution of those which are available. Consequently substantial unemployment of human resources exists and this is concentrated on those groups who are socially disadvantaged and who lack industrial and political power. The concentration of such groups in the lower segments of the labour market lays them open to further exploitation in that they receive lower pay relative to their productivity than those more fortunately placed.

This structuring based on inequality of job opportunity is socially reinforced because jobs are classified as skilled more by the social characteristics of the incumbents than the content of the job. This hierarchy is further sanctified by the theorizing (or perhaps theologising) of orthodox economists who equate high wages with labour 'quality'. This categorization is socially useful in that it allows the targeting of the cost of economic adversity on those groups without social and political power, a bonus which is multiplied when the manifestation of the increased deprivation provides readily acceptable explanations for unemployment. But the proponents of supply side explanations of low pay and unemployment should ponder on the fact that between 1938 and 1942, after two decades of very high unemployment, employment in Britain increased by 3 million and unemployment virtually disappeared. Of the total in employment 3.5 million prime age workers were diverted into the armed forces and their place in production was taken by previously domestically employed women and the unemployed – many regarded hitherto as unemployable. What made the difference was not a massive training programme or other 'quality'

raising exercises, there was no time for that, the transformation came because war needs made jobs available.

See Also

- ▶ Minimum Wages
- ▶ Segmented Labour Markets

Bibliography

Craig, C., J. Rubery, R. Tarling, and F. Wilkinson. 1985. Economic, social and political factors in the operation of labour markets. In *New approaches to economic life*, ed. B. Roberts, R. Finnegan, and D. Gallie. Manchester: Manchester University Press.

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

Layard, R. 1986. *How to beat unemployment*. Oxford: Oxford University Press.

Marshall, A. 1920. *Principles of economics*, 8th ed. London: Macmillan. Reprinted, 1952.

Rubery, J. 1978. Structural labour markets. Worker organisation and low pay. *Cambridge Journal of Economics* 2(1): 17–36.

Spence, M. 1973. Job market signalling. *Quarterly Journal of Economics* 87(3): 355–374.

Stedman Jones, G. 1984. *Outcast London*. London: Penguin.

Turner, H.A. 1962. *Trade unions growth, structure and policy*. London: Allen & Unwin.

Wilkinson, F. (ed.). 1981. *Dynamics of labour market segmentation*. London: Academic.

Lowe, Adolph (1893–1995)

Edward J. Nell

Keywords

Economic growth; Freedom; German hyperinflation; Instrumental analysis; Lowe, A.; Planning; Technical change

JEL Classifications

B31

Born on 4 March 1893 in Stuttgart, Adolph Lowe was educated at Berlin and Tübingen and received the Dr. Juris. from Tübingen in 1918. From 1919 to 1924 he was Section Head in the Ministries of Labour and Economics of the Weimar Republic, and was largely responsible for the practical planning and management of the currency reforms that brought the great hyperinflation to an end. From 1924 to 1926 he was Head of the International Division of the Federal Statistical Bureau, a politically sensitive post in the light of disputes over reparations payments. In 1926 he became Director of Research at the Institute of World Economics at the University of Kiel, where he established an important centre for research into business cycles and their control and regulation through planning. In 1931 he was appointed Professor of Political Economy at the University of Frankfurt, where he joined the leaders of a major renaissance in social and socialist thinking. But in March 1933 he became the first professor in the social sciences to be fired by Hitler. He moved immediately to England, where he held a post at Manchester until 1940, when he moved to the New School for Social Research in New York, where he was Professor of Economics, Director of Research at the Institute of World Affairs, and then Professor Emeritus, remaining active in the Department until his return to Germany, in March 1983, 50 years after his forced departure. In 1984 he was awarded the Dr. *honoris causa* by the University of Bremen.

His publications include 'Wie Ist Konjunkturtheorie Überhaupt Moglich?' (1926), *Economics and Sociology* (1935), *The Price of Liberty* (1937), 'The Classical Theory of Economic Growth' (1954), *On Economic Knowledge* (1965, 1977) and *The Path of Economic Growth* (1976). *Economic Means and Social Ends*, edited by Robert L. Heilbroner, was published in 1969 in honour of Professor Lowe's 75th birthday.

Unlike many economists, Lowe considered economics inseparable from social inquiry in general. In his view, the central question of economics is the determination of the path of economic growth and its relation to technical progress and social change. Lowe developed a strikingly simple three-sector model in which structural changes

during expansion could be displayed. Growth will normally not take place in a balanced manner; more commonly the actual path will be a 'traverse' from one desired path to another, which is likely to shift again before it is reached. But the problem has to be understood in the light of what Lowe calls 'instrumental analysis'. Conventional economic theory begins with knowledge of the prevailing situation and a set of well-defined behavioural laws, based on maximizing. From these two givens one can deduce/predict the future configuration of the economy. This approach worked well in the early stages of capitalism, when the pressure of poverty on labour and competition on capital ensured stable patterns of behaviour. But mass production and economies of scale undermine competition, while affluence and unionization, together with the growth of the middle class, lead both to unpredictable wage bargaining and to unstable consumer spending. Tastes become volatile, while consumption can be postponed or redirected, and businesses plan strategically, often in cooperation with their rivals, instead of maximizing on a short horizon – so the traditional approach is no longer appropriate. The historical conditions do not constrain behaviour sufficiently for maximizing models, even complex ones, to picture it accurately, so that the conventional method must be set aside. (Which means, as well, that the forces of the market cannot be relied upon; they are no longer determinate.) Instead, the givens should be the existing conditions and the *desired terminal position*, and the job of economic analysis then becomes to find the 'goal-adequate' sequences of change, together with the stimuli and/or constraints that will create the necessary behaviour patterns. Such stimuli and constraints must be imposed by government. Economic analysis becomes a form of planning, and Lowe's work in his last years analysed the relation of planning to freedom.

Selected Works

1926. Wie ist Konjunkturtheorie überhaupt möglich? *Weltwirtschaftliches Archiv* 24 (2): 165–197.

1935. *Economies and sociology: A plea for co-operation in the social sciences*. London: G. Allen & Unwin.

1937. *The price of liberty: A German on contemporary Britain*. London: L. and Virginia Woolf at the Hogarth Press.

1954. The classical theory of economic growth. *Social Research* 21: 127–158.

1965. *On economic knowledge*. New York: Harpers; London: Longmans. Enlarged ed. New York/London: M.E. Sharpe, 1977.

1969. *Economic means and social ends: Essays in politics economics*, ed. R.L. Heilbroner. Englewood Cliffs: Prentice-Hall.

1976. *The path of economic growth*. Cambridge: Cambridge University Press. 1988. *Has freedom a future?* New York: Praeger.

Low-Income Housing Policy

Edgar O. Olsen

Abstract

Low-income housing assistance is an important part of the welfare system in many countries. This article discusses the rationale for this government activity, describes the most important differences between different low-income housing programmes, explains why economic theory has limited implications for the effects of these programmes, and summarizes the evidence on their most important effects. The most important finding of the empirical literature on the effects of different housing programmes from the viewpoint of housing policy is that recipient-based housing assistance has provided equally good housing at a much lower total cost than any type of unit-based assistance.

Keywords

Crowding out; Low-Income housing policy; Neighbourhood effects; Public housing;

Recipient-Based housing assistance; Residential segregation; Unit-Based housing assistance

JEL Classifications
H5

Low-income housing assistance is an important part of the welfare system in many countries.

Rationales

The most compelling rationale for this government activity is that some taxpayers care about low-income households and think that the decision makers in some of these households spend too little of their income on housing for their own good. Another important argument is that some taxpayers are particularly concerned about the well-being of the children in low-income households and prefer housing subsidies to unrestricted cash grants in order to better target assistance to the objects of their concern. These rationales imply that a successful housing programme induces its recipients to occupy better housing and consume less of other goods than they would choose in response to an unrestricted cash grant in an amount equal to the housing subsidy.

Programme Types

Governments have tried many methods of providing housing assistance. The most important distinction between rental housing programmes is whether the subsidy is attached to the dwelling unit or to the assisted household. If the subsidy is attached to a rental dwelling unit, each family must accept the particular unit offered in order to receive assistance and loses its subsidy when it moves. Each family offered recipient-based rental assistance has a choice among many units in the private market that meet the programme's standards, and the family can retain its subsidy when it moves. The analogous distinction for homeownership programmes is between programmes that require eligible families to buy from selected sellers in order to receive a subsidy and programmes that provide subsidies to eligible families that are free to buy from any seller that provides housing meeting the programme's standards.

There are two broad types of unit-based rental assistance, namely, public housing and privately owned subsidized projects. Public housing projects are owned and operated by government entities. In public housing programmes, civil servants make all of the decisions made by private owners of unsubsidized housing. Governments also contract with private parties to provide unit-based assistance in subsidized housing projects. In the United States, the majority of these private parties are for-profit firms, but non-profit organizations have a significant presence. Under most programmes, these private parties agree to provide rental housing meeting certain standards at restricted rents to households with particular characteristics for a specified number of years. The overwhelming majority of the projects were newly built under a subsidized construction programme. Almost all of the rest were substantially rehabilitated as a condition for participation in the programme. None of the programmes that subsidize privately owned projects provide subsidies to all suppliers who would like to participate.

In 2004, the United States government spent about $15 billion on its housing voucher programme, more than $15 billion to subsidize private projects for low-income households, and about $7.5 billion to subsidize public housing projects. The US Department of Housing and Urban Development's Section 8 New Construction and Substantial Rehabilitation Program and the Internal Revenue Service's Low-Income Housing Tax Credit Program are the two largest programmes that subsidize private rental projects, accounting for about 75 per cent of public expenditure on programmes of this type. In total, these rental programmes served about seven million households. During the same year, the US government spent only $4 billion to subsidize low-income homeowners. These programmes tend to provide shallower subsidies to households with substantially higher incomes than the rental programmes.

Theory

Economic theory that accounts for the most rudimentary features of real housing programmes does not have strong implications about their effects. For example, these programmes may induce households to occupy worse housing even if housing is a normal good. Such counter-intuitive outcomes result from the nonlinear budget frontiers facing households offered housing assistance. For instance, a household offered a unit in a subsidized housing project is offered an all-or-nothing choice of a particular dwelling unit at a below-market rent. This unit might be worse than the household's current unit, but the household may accept the offer because the reduction in its rent enables it to consume more of other goods.

Evidence

The remainder of this article summarizes the evidence on the effects of the major rental housing programmes in the United States. The United States has rental programmes of each broad type, and a disproportionate share of the evidence on the performance of low-income housing programmes throughout the world pertains to these programmes. Homeownership programmes are a small part of the current system, and little is known about their effects.

Different rental housing programmes have different effects. Indeed, the same programme has different effects in different circumstances. Olsen (2003) provides a more detailed account of the evidence on the performance of individual programmes, and the bibliography to this article contains references to some of the more important recent studies. This article endeavours to characterize what is typical of these programmes and the differences in the average effect of programmes of different types.

The most important finding of the empirical literature on the effects of different housing programmes from the viewpoint of housing policy is that recipient-based housing assistance has provided equally good housing at a much lower total cost than any type of unit-based assistance. The reasons for this result suggest that it would apply generally. These reasons include the absence of a financial incentive for good decisions on the part of civil servants who operate public housing, the excessive profits that inevitably result from allocating subsidies to selected developers of private subsidized projects, and the distortions in usage of inputs resulting from the subsidy formulas. Another reason for the excess cost of unit-based assistance is that this assistance is usually tied to the construction of new units. The least expensive approach to improving the housing conditions of low-income households involves heavy reliance on upgrading the existing housing stock.

Since housing programmes are intended to produce particular changes in consumption of housing services compared with consumption of other goods, knowledge of these changes is important for evaluating these programmes. The overwhelming majority of recipients of housing assistance occupy better housing than they would occupy in the absence of assistance. More importantly, they typically occupy better housing than they would occupy if they were given cash grants in amounts equal to their housing subsidies. Most recipients of rental housing assistance pay significantly less for their housing and hence have more to spend on other goods.

One aspect of the housing bundle broadly conceived that has attracted considerable attention is its neighbourhood. Recipients of tenant-based vouchers and occupants of privately owned subsidized projects typically live in somewhat better and less racially segregated neighbourhoods than in the absence of housing assistance. Occupants of public housing typically live in noticeably worse and more racially segregated neighbourhoods.

A careful theoretical analysis that accounts for a key feature of low-income housing programmes has shown that, even if the subsidy under the programme declines with increases in earnings and leisure is a normal good, the programme will not necessarily induce the recipient to work less (Schone 1992). Nevertheless, evidence based on a controlled experiment indicates that voucher recipients reduce their earnings about 13 per cent on average (Patterson et al. 2004). Other evidence indicates that programmes of unit-based

assistance have somewhat larger work disincentive effects (Olsen et al. 2005).

Low-income housing programmes differ substantially from unrestricted cash grants in their effects. The mean value of project-based housing assistance as judged by recipients is much less than 75 per cent of the mean housing subsidy (that is, the difference between the market rent of the subsidized unit and the tenant's contribution). The mean value of tenant-based housing assistance as judged by recipients is about 80 per cent of the mean housing subsidy.

Consistent with their intentions, the mean benefit to recipients in these programmes is greater for poorer and larger households among households that are the same in other respects. Mean benefit varies little with the age, race and sex of the head of the household after other household characteristics are accounted for. The variance in benefit among recipients with the same characteristics is large under construction programmes that have produced new units for many years. In these mature construction programmes, there is an enormous difference between the best and the worst units, and a tenant with specified characteristics would pay the same rent for these units.

Unit-based or recipient-based housing programmes can make the neighbourhoods into which subsidized households move better or worse places to live. Neighbourhood property values capture these effects. On average across all units in a programme, the evidence indicates that no programme has had a significant effect on neighbourhood property values.

Housing programmes affect the rents of unsubsidized units with unchanging characteristics. Evidence from the Housing Assistance Supply Experiment indicates that an entitlement housing voucher programme for which the poorest 20 per cent of the population is eligible will have small effects on market rents (Lowry 1983). No evidence is available for construction programmes. However, economic theory suggests that, if a construction programme leads to a larger housing stock, it will result in higher market rents because it will drive up the prices of inputs used heavily in the housing industry. This effect might be small, however, because the evidence indicates that subsidized construction crowds out unsubsidized construction to a considerable extent (Malpezzi and Vandell 2002; Sinai and Waldfogel 2005; and references in Olsen 2003).

An important recent literature estimates a wide range of impacts of offering portable vouchers to families living in the worst public housing projects or in public housing projects in the poorest neighbourhoods. The larger strand of this research is based on data from a controlled experiment called Moving to Opportunity, in which one experimental group was offered a housing voucher without any restriction on the neighbourhood where it could be used and another experimental group had to move for at least a year to a neighbourhood where the poverty rate was less than ten per cent prior to the experiment (Orr et al. 2003). These treatments led their recipients to live in better housing and neighbourhoods without a reduction in expenditure on other goods. However, they did not lead to some expected outcomes. After four to seven years in the experiment, the treatment groups did not increase their earnings and their children's educational performance did not improve. With a few notable exceptions such as the mental health of girls and their mothers, the treatments had minimal effects on health outcomes. The treatments generally had effects in opposite directions on the delinquency and risky behaviour of boys and girls. The effects on boys were negative, though these effects were not usually statistically significant. A smaller strand of this literature is based on data on natural experiments such as when public housing tenants must move because their project is torn down (Jacob 2004).

See Also

- ▶ Crowding Out
- ▶ Housing Policy in the United States
- ▶ Housing Supply
- ▶ Welfare State

Bibliography

Jacob, B. 2004. Public housing, housing vouchers, and student achievement: Evidence from public housing demolitions in Chicago. *American Economic Review* 94: 233–258.

Lowry, I.S., ed. 1983. *Experimenting with housing allowances: The final report of the housing assistance supply experiment*. Cambridge, MA: Oelgeschlager, Gunn & Hain.

Malpezzi, S., and K. Vandell. 2002. Does the low-income housing tax credit increase the supply of housing? *Journal of Housing Economics* 11: 360–380.

Olsen, E.O. 2003. Housing programs for low-income households. In *Means-tested transfer programs in the United States*, ed. R. Moffitt. Chicago: University of Chicago Press.

Olsen, E.O., C.A. Tyler, J.W. King, and P.E. Carrillo. 2005. The effects of different types of housing assistance on earnings and employment. *Cityscape* 8: 163–187.

Orr, L., et al. 2003. *Moving to opportunity for fair housing demonstration program: Interim impacts evaluation*. Washington, DC: US Department of Housing and Urban Development.

Patterson, R., et al. 2004. *Evaluation of the welfare to work voucher program: Report to congress*. Washington, DC: US Department of Housing and Urban Development.

Schone, B.S. 1992. Do means tested transfers reduce labor supply? *Economics Letters* 40: 353–358.

Sinai, T., and J. Waldfogel. 2005. Do low-income housing subsidies increase the occupied housing stock? *Journal of Public Economics* 11–12: 2137–2164.

Lucas Critique

Lars Ljungqvist

Abstract

The 'Lucas critique' is a criticism of econometric policy evaluation procedures that fail to recognize that optimal decision rules of economic agents vary systematically with changes in policy. In particular, it criticizes using estimated statistical relationships from past data to forecast the effects of adopting a new policy, because the estimated regression coefficients are not invariant but will change along with agents' decision rules in response to a new policy. A classic example of this fallacy was the erroneous inference that a regression of inflation on unemployment (the Phillips curve) represented a structural trade-off for policy to exploit.

Keywords

Expectations; Lucas Critique; Macroeconomic policy evaluation; Optimization behaviour; Phillips curve; Rational expectations; Rational expectations econometrics; Real vs. nominal shocks

JEL Classifications

D4; D10

The 'Lucas Critique' is a criticism of econometric policy evaluation procedures that fail to recognize the following economic logic:

> [G]iven that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any changes in policy will systematically alter the structure of econometric models. (Lucas 1976, p. 41)

At the time of his writing, Robert E. Lucas (1976) was criticizing the prevailing approach to quantitative macroeconomic policy evaluation for ignoring this logic and, hence, as being fundamentally inconsistent with economic theory. To fully appreciate Lucas's critique, we first consider a general theoretical argument and then turn to a particular example.

At each date t there is a vector s_t of state variables summarizing all aspects of the history that are relevant to the economy's future evolution; for example, the vector might include the economy's capital stock. The economy is also described by a vector x_t of government policy variables and a vector ε_t of random shocks – for example, shocks to technology or to government policy. For given specifications of the processes governing x_t and ε_t, it is common in macroeconomic theory to analyse models that yield an equilibrium law of motion in form of a difference equation,

$$s_{t+1} = f(s_t, x_t, \varepsilon_t). \tag{1}$$

(For many textbook examples of stochastic rational expectations models that yield such a recursive equilibrium representation; see Ljungqvist and Sargent 2004). Equation (1) is also the point of departure for the econometric policy evaluation procedures criticized by Lucas, who argued that their approach failed to recognize the optimization behaviour of economic agents that is implicit in Eq. (1). Specifically, the criticized approach proceeds as follows. First, historical data are used to estimate the equation

$$s_{t+1} = F(\theta, s_t, x_t, \mu_t), \tag{2}$$

where F is specified in advance, θ is a fixed parameter vector to be estimated, and μ_t is a vector of random disturbances. Second, with the use of the estimated Eq. (2), policy evaluations are performed by comparing economic outcomes for different paths of government policy variables $\{x_t\}$. The policy choice that produces the most desirable economic outcome is deemed to be the best policy. But, as argued by Lucas, this approach violates the premises for economic theory because the parameter vector θ depends partly on agents' decision rules that are not invariant to the conduct of government policy. That is, if the government changes its policy, the parameter θ will also change, so that the consequences of a new policy cannot be evaluated on the basis of the historical relationship in Eq. (2).

Lucas's argument is best illustrated with an example. Consider the classic example of the so-called 'Phillips curve'. Phillips (1958) had estimated a negative relationship between wage inflation and unemployment using British data for the period 1861–1957. Samuelson and Solow (1960) and others interpreted this and related empirical findings as evidence of a structural trade-off between an economy's inflation rate and its unemployment rate. That is, the parameter θ in Eq. (2), estimated with historical data, was considered to be fixed and to describe how unemployment would respond to inflation outcomes associated with different monetary policies. Friedman (1968) and Phelps (1968) argued against the existence of such an exploitable trade-off because it was inconsistent with economic theory based on rational agents. To understand the fallacy of the Phillips curve and its extension – the fallacy of the econometric policy evaluation procedures criticized by Lucas – consider the monetary model of Lucas (1972). Exchange in the economy takes place in physically separated markets. Producers in a market base their output decisions on the local market-clearing price level without knowing the current economy-wide price level. The price in a market varies stochastically because there are exogenous random shocks both to the distribution of producers across markets and to the aggregate quantity of nominal money, none of which is directly observable to the agents. Hence, information on the current state of these real and monetary shocks is transmitted to agents only through the price in the market where each agent happens to be. In an equilibrium, producers in a market would like to increase their output in response to a high price driven by real but not nominal shocks. A high price due to a real shock means that the ratio of producers to consumers is low in that market and, therefore, profits on sales are high in real terms (when evaluated in terms of the economy-wide price level). But a high price in a market due to an expansion of the aggregate quantity of nominal money means that prices tend to be high in all markets and, therefore, profits on sales are high in nominal but not real terms. The inference and decision problems solved by the agents in this model are shown to give rise to a Phillips curve, as had been estimated with real-world data, but where the model's apparent trade-off between inflation and output cannot be systematically exploited by the government in its choice of monetary policy.

To further convey the insights from this general equilibrium model of the Phillips curve, we adopt a version of Lucas's (1976) simplified model that does not spell out all the details of the economic environment but instead postulates three equations that capture the forces at work in the fully articulated model. The economy-wide price level (in logs), p_t, is given by

$$p_t = \overline{p}_t + m_t, \tag{3}$$

where $\overline{p}_t$ reflects a systematic component of monetary policy that is known to all agents, and m_t reflects an i.i.d. shock to monetary policy. It is assumed that the random variable m_t is normally distributed with mean zero and variance σ_m^2. The price (in logs) in market i at time t, p_{it}, is given by

$$p_{it} = p_t + z_{it}, \tag{4}$$

where z_{it} is a deviation from the economy-wide price level because of shocks to the distribution of producers across markets. The real shock z_{it} is assumed to be a normal, i.i.d. random variable with mean zero and variance σ_z^2. Finally, let y_{it} denote the log-deviation of output from its 'natural rate' in market i at time t which varies with the perceived, relative price:

$$y_{it} = \alpha[p_{it} - E(p_t|I_{it})], \tag{5}$$

where $\alpha > 0$ reflects intertemporal substitution possibilities in supply (determined by technological factors and tastes for substituting labour over time), and $E(\,|I_{it})$ denotes the mathematical expectation conditioned upon information I_{it} available in market i at time t. The agents' prediction problem in Eq. (5) is straightforward to solve (see, for example, Ljungqvist and Sargent 2004, ch. 5):

$$\begin{aligned} E(p_t|I_{it}) &= E(p_t|p_{it},\overline{p}_t) \\ &= (1-\Omega)p_{it} + \Omega\overline{p}_t, \end{aligned} \tag{6}$$

where $\Omega = \sigma_z^2/(\sigma_m^2 + \sigma_z^2)$. The substitution of Eqs. (3), (4) and (6) into Eq. (5) yields

$$y_{it} = \alpha\Omega(m_t + z_{it}). \tag{7}$$

Thus, output in market i varies with the sum of nominal and real shocks, $(m_t + z_{it})$, because producers cannot perfectly disentangle these shocks but must make inferences based on the observed price p_{it}. Producers' willingness to vary output from its natural rate depends on how likely observed price variations are due to real rather than nominal shocks, as captured by the magnitude of $\Omega \in [0, 1]$. Under the assumption of a large number N of markets, the real shocks, $\{z_{it}\}$, cancel each other out when averaged over markets, and the economy's deviation from its natural rate of output, y_t, becomes

$$y_t = \frac{1}{N}\sum_{i=1}^{N} y_{it} = \alpha\Omega m_t = \alpha\Omega(p_t - \overline{p}_t), \tag{8}$$

where the last equality invokes Eq. (3) and, hence, the economy exhibits a positive relationship between unanticipated inflation and output.

If estimations were performed using data on output and inflation from the described economy, we would find a Phillips curve along which increases in inflation are associated with higher output realizations. However, any attempts by the government to exploit that relationship would fail. For example, a government that permanently increases the growth rate of the money supply to generate higher inflation in order to stimulate output will ultimately see no real effects from that change in policy. The reason for this is that, after agents have become aware of the higher underlying inflation rate in the economy, they will change their expectations when making predictions about relative price movements due to real disturbances. Formally, the change in monetary policy represents an increase in the component $\overline{p}_t$ and, when that systematic change becomes known to the agents, it will not affect unanticipated inflation, $(p_t - \overline{p}_t) = m_t$, so output is left unaffected in Eq. (8).

This example illustrates Lucas's general criticism of econometric policy evaluation procedures that fail to recognize that the estimated Eq. (2) depends partly on agents' decision rules and is therefore not invariant to changes in government policy. For a proper policy evaluation procedure, we need to revise the econometric formulation in Eq. (2) so that it becomes consistent with equilibrium outcomes as represented by Eq. (1). Recall that the latter equation is derived for given specifications of the processes governing x_t and ε_t. In particular, to analyse agents' optimization behaviour, we need to specify the environment in which they live, including their perceptions about future government policy. As Lucas (1976, p. 40) remarked, 'one cannot meaningfully discuss optimal decisions of agents under arbitrary sequences $\{x_t\}$ of future shocks'. Instead, Lucas suggested

that one proceeds by viewing government policy as a function of the state of the economy,

$$x_t = G(\lambda, s_t, \eta_t), \tag{9}$$

where λ is a parameter vector that characterizes government policy, and η_t is a vector of random disturbances. Then the new version of Eq. (2) becomes

$$s_{t+1} = F(\theta(\lambda), s_t, x_t, \mu_t), \tag{10}$$

and the econometric problem is that of estimating the function $\theta(\lambda)$. A change in government policy is viewed as a change in the parameter λ affecting the behaviour of the system in two ways: first, by altering the time series behaviour of $\{x_t\}$, and second, by leading to modification of the parameter θ governing the rest of the system, which reflects changes in agents' decision rules in response to the new policy.

A constructive response to the Lucas critique has been the development of rational expectations econometrics. A goal of that approach has been to estimate the 'primitives' of dynamic rational expectations models, in the form of parameters describing tastes and technologies. If historical data can be used to obtain such estimates, the economic model can in principle be used to evaluate alternative government policies that could be without precedent, as explained by Lucas and Sargent (1981). That is, knowledge about the primitives of a model enables us to derive agents' decision rules and equilibrium outcomes for any specified policy process. In terms of Eq. (10), this explains how the function $\theta(\lambda)$ could conceivably be estimated even if the historical data have been generated under a single government policy λ.

Though one of the key contributors to the methodology of rational expectations econometrics, Sargent (1984) has raised a philosophical conundrum with this approach to policy evaluation (as earlier discussed by Sargent and Wallace 1976). Suppose that the primitives of an economic model have been estimated during an estimation period in which government policy was specified to be λ, and then the estimated model is used to compare alternative policies in order to find the best future policy λ^*. But such a procedure leads to an internal contradiction under the assumption of rational expectations, because, if the procedure were in fact likely to be persuasive in having the policy recommendation actually adopted soon, it would mean that the original econometric model with it specified policy λ had been misspecified. As pointed out by Sargent (1984, p. 413): 'A rational expectations model during the estimation period ought to reflect the procedure by which policy is thought later to be influenced, for agents are posited to be speculating about government decisions into the indefinite future.'

Given its fundamental impact on questions of economic policy both in practice and in theory, the Lucas critique figured prominently in the list of contributions when the Royal Swedish Academy of Sciences (1995) awarded Robert E. Lucas, Jr. the Nobel Prize in economics 'for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy.'

See Also

- Phillips Curve
- Rational Expectations

Bibliography

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Ljungqvist, L., and T.J. Sargent. 2004. *Recursive macroeconomic theory*. 2nd ed. Cambridge, MA: MIT Press.

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

Lucas, R.E., Jr. 1976. Econometric policy evaluation: A critique. In *The phillips curve and the labor market*, ed. K. Brunner and A. Meltzer, Vol. 1 of Carnegie-Rochester conference on public policy. Amsterdam, North-Holland.

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

Phelps, E.S. 1968. Money wage dynamics and labor market equilibrium. *Journal of Political Economy* 76: 687–711.

Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the

United Kingdom, 1861–1957. *Econometrica* 25: 283–299.
Royal Swedish Academy of Sciences 1995. The Sveriges Riksbank prize in economic sciences in memory of Alfred Nobel for 1995. Press release, 10 October. Online. Available at http://nobelprize.org/economics/laureates/1995/press.html. Accessed 4 Oct 2006.
Samuelson, P.A., and R.M. Solow. 1960. Analytical aspects of anti-inflation policy. *American Economic Review* 50: 177–194.
Sargent, T.J. 1984. Autoregressions, expectations, and advice. *American Economic Review* 74: 408–415.
Sargent, T.J., and N. Wallace. 1976. Rational expectations and the theory of economic policy. *Journal of Monetary Economics* 2: 169–183.

Lucas, Robert (Born 1937)

Levon Barseghyan

Abstract

Robert E. Lucas, Jr is one of the most influential economists of our time. His work on rational expectations offered a truly new way of thinking about economics and policy that led to most of the recent successes in macroeconomics. Lucas's path breaking research on so many issues of vital importance has advanced the frontier of science and set the stage for new exciting discoveries.

Keywords

Business cycles; Cash-in-advance constraint; Dynamic stochastic general equilibrium models; Economic growth; Endogenous growth theory; Human capital; Inflation; Lucas Critique; Lucas trees; Lucas, R; Monetary business cycle models; Monetary shocks; Money growth; Neoclassical growth theory; Neutrality of money; Overlapping generations models; Phillips curve; rational expectations; Time consistency of monetary and fiscal policy

JEL Classification

B31

In 1995, Robert E. Lucas, Jr received the Nobel Prize in Economic Sciences 'for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy' (Press Release announcing the Nobel Prize 1995; repr. in Svensson 1996, p. 1).

Robert Lucas was born in Yakima, Washington on 15 September 1937. He received his BA in History in 1959, and his Ph.D. in Economics in 1964, both from the University of Chicago. He began his career as an assistant professor at Carnegie Mellon University, where he became an associate professor in 1967 and a full professor in 1970. He joined the Department of Economics at the University of Chicago as a full professor in 1975, and since 1980 has served as John Dewey Distinguished Service Professor of Economics there. He is a fellow of the Econometric Society, the American Academy of Arts and Sciences, and the American Finance Association; a member of the National Academy of Sciences and the American Philosophical Society and a titular member of the European Academy of Arts, Sciences and Humanities. Lucas served as the President of the Econometric Society in 1997 and as the President of the American Economic Association in 2002.

Robert Lucas's seminal contributions in the early 1970s led to a paradigm shift in macroeconomics: the rational expectations revolution. By the late 1970s–early 1980s, due to the efforts of Robert Lucas and others (including Robert Barro, William Brock, Edward Prescott, Thomas Sargent and Neil Wallace) the frontier of macroeconomic research had moved away from models with static or adaptive expectations towards models in which agents act in their best interest, utilizing all available information about past, present and future. As a result, dynamic stochastic general equilibrium models with rigorous microfoundations have been developed to understand economic fluctuations and growth and to analyse the effects of monetary and fiscal policies. While these models have become increasingly complex in an effort to better understand the economy, almost all of them are built on the principles set forth by Robert Lucas.

The Beginning of the Rational Expectations Revolution: Expectations and the Neutrality of Money

Robert Lucas's 'Expectations and Neutrality of Money', published in 1972 in the *Journal of Economic Theory*, was the first paper to incorporate the idea of rational expectations into a dynamic general equilibrium model. (rational expectations were introduced by Muth 1961. In their groundbreaking study of investment under uncertainty, Lucas and Prescott 1971 applied the notion of rational expectations in a dynamic partial equilibrium model of a competitive industry facing stochastic demand.)

The agents in Lucas's (1972b) model are fully rational: based on the available information, they form expectations about future prices and quantities, and based on these expectations they act to maximize their expected lifetime utility. This paper also was the first to provide sound theoretical underpinnings to Milton Friedman's (1968) and Edmund Phelps's (1968) view of the long-run neutrality of money, and at the same time to provide an explanation of the observed positive correlation between output and inflation, famously depicted by the Phillips curve.

Lucas's model is built on Paul Samuelson's (1958) overlapping generations model. Agents live for two periods. In each period the young generation works, consumes and saves. The old generation consumes its savings. Goods are perishable and there is only one savings instrument in the economy, money.

The population in the economy is allocated into two distinct markets (islands) across which no communication is possible. The old generation is equally divided between the islands. The allocation of the young generation across the islands is a random variable. The amount of money holdings by the old generation is also a random variable, because it depends on the realization of a random shock in the money growth rate: each dollar carried from one period to another is multiplied by the realized money growth rate between these two periods. Agents do not observe the current allocation of young across islands and the money growth rate, but know their underlying probability distributions. To solve for the optimal amount of labour supply and savings, the young must form expectations about the future value of money, that is, the future price level. How does one form such expectations? Lucas's answer is, rationally. He defined and explicitly solved for the rational expectations equilibrium, in which agents correctly predict how the price level depends on the state of the economy. Of course, to do so each agent also must correctly understand the actions of all other agents in the current and future generations and how these actions affect prices (and quantities).

In the model, the positive correlation between the money growth rate and output arises because the young, when faced with a high demand for their goods, are unable to distinguish its source: the demand could be high because of a higher money growth rate, or because of a lower fraction of the young workers on the island. Due to their inability to infer exactly the source of the high demand, the young find it optimal to produce whenever they face a high demand. Consequently, a positive money growth shock leads to an economic expansion on both islands. Without uncertainty about the money growth rate, the neutrality of money is immediately attained. Any pre-announced proportional money growth rule – for example, the $k\%$ rule advocated by Milton Friedman – results in the same real outcomes.

Lucas showed that invariance of real outcomes to the pre-announced part of the money growth rule holds also when there are shocks to money growth. This finding is often characterized as a 'policy ineffectiveness' result, because it implies that, although there is a positive correlation between output and money growth, this correlation cannot be exploited by the monetary authority to influence real economic activity.

Prior to Lucas's (1972b) work, economists often emphasized that a distinction should be drawn between the long-run and the short-run effects of monetary shocks. An important corollary of Lucas's work is that this distinction often is misleading. The true distinction must be made between anticipated and unanticipated monetary disturbances, because their effects on real

economic activity are likely to be very different. Most of the subsequent monetary business cycle literature embraces this distinction.

Econometric Policy Evaluation: The Lucas Critique

Lucas (1976), known as the Lucas critique, marked the turning point in how economists approached econometric policy evaluation. Thomas Sargent's (1996) account of the events following the Lucas critique gives a sense of its tremendous impact:

> [W]e didn't understand what was going on until, upon reading Lucas's 'Econometric Policy Evaluation' in Spring of 1973, we were stunned into terminating our long standing Minneapolis Fed research project to design, estimate and optimally control a Keynesian macroeconometric model. We realized that Kareken et al. (1973) defense of the 'look-at-everything' feedback rule for policy – which was thoroughly based on 'best responses' for the monetary authority exploiting a 'no response' private sector – could not be the foundation of a sensible research program, but was better viewed as a memorial plaque to the Keynesian tradition in which we had been trained to work. (Sargent 1996, p. 539)

The essence of the Lucas critique stems naturally from the concept of rational expectations. Indeed, rationality of the private sector implies that it cannot be modelled as a 'no response' entity. Rather, any observed or anticipated change in monetary policy, including the 'best response' of the monetary authority, will induce the 'best responses' from the agents in the private sector. This, in turn, implies that the effects of a new policy cannot be assessed according to econometric estimation of the private sector's behaviour under the old policy.

Other Major Contributions

Robert Lucas has made several other major contributions in different areas of economics. A small subset of them is presented below, in chronological order.

Lucas (1978a) elegantly introduced the first general equilibrium model of asset pricing. In the model economy, physical assets are represented by what nowadays typically is referred to as 'Lucas trees': infinitely lived objects that generate stochastic dividends (fruits). Lucas explicitly derived asset prices as functions of the economy's state variables. The logic of Lucas's asset pricing equation forms the foundation of many models in macro and financial economics.

Lucas (1980b) and Lucas and Stokey (1987) helped to lay the foundations of monetary economics. The ideas and the methodology developed in these papers continue to guide monetary economists, particularly in applied research. Lucas (1980b) is the first general equilibrium study of the determination of prices in an economy in which the use of money arises from a cash-in-advance constraint. The model in Lucas and Stokey (1987), which is the prototype for a number of widely used dynamic stochastic general equilibrium monetary models, features both real and nominal shocks. Methods developed by Lucas and Stokey for establishing the existence of, characterizing and solving for the equilibrium of such models have proven to be powerful tools in applied and theoretical research.

Lucas (1982) extended the logic of his earlier contributions, Lucas (1978a) and Lucas (1980b), to a two-country stochastic general equilibrium model with infinitely lived agents, in which he explicitly derived formulas for pricing real assets and nominal bonds as well as for determining exchange rates. The framework developed in this paper serves as a point of departure for many models in international economics.

Lucas and Stokey (1983) is a major contribution to modern public finance. Lucas and Stokey studied the Ramsey (1927) problem – the problem of optimal taxation when non-distortionary tax instruments are unavailable – in dynamic stochastic economies without physical capital. Their paper provided a number of important insights about the structure and time consistency of optimal fiscal and monetary policies. Lucas and Stokey showed that a sufficiently rich debt maturity structure could allow for time consistency of the optimal fiscal policy.

Lucas (1988) is a seminal contribution in the economic development and growth literature (see also the 1991 Fisher and Shultz Lecture at the

European Meetings of Econometric Society, published as Lucas 1993). Lucas (1988) and an earlier paper by Paul Romer (1986) heralded the birth of endogenous growth theory and the resurgence of research on economic growth in the late 1980s and the 1990s. These papers offered an escape from 'the straightjacket of the neoclassical growth model, in which the long term per capita (output) growth is pegged by the rate of exogenous technological progress' (Barro and Sala-i-Martin 2004, p. 19), by showing that factor accumulation does not need to run into diminishing returns to scale and, therefore, could lead to perpetual growth. In particular, Lucas emphasized the role of human capital, and externalities generated by it, as important sources of long-run economic growth.

Robert Lucas has written a number of seminal books. Among them are *Models of Business Cycles* (1987) and, with N. Stokey and E. Prescott, *Recursive Methods in Economic Dynamics* (1989). The former presents a critical assessment of the business cycle literature of the 1970s and the early 1980s and offers novel insights about economic fluctuations. This monograph contains Lucas's famous calculation of the cost of business cycles, which he argued to be insignificant. (In a similar spirit, Lucas 2000a provided a quantitative assessment of the welfare cost of inflation. In this paper, he found that the gains from reducing inflation could be non-negligible. Subsequent research often has taken his calculations of the cost of business cycles and of the cost of inflation as benchmarks.) Another indispensable volume, *Recursive Methods in Economic Dynamics* (1989), deals with stochastic dynamic programming. It has been widely used as a textbook in graduate macroeconomics courses and as a guide for formulating and solving dynamic stochastic general equilibrium models.

See Also

- ▶ Lucas Critique
- ▶ Monetary Business Cycles (Imperfect Information)
- ▶ Neutrality of Money
- ▶ Phillips Curve
- ▶ Rational Expectations

In writing this article I have drawn from Fisher (1996), *Hall* (1996), *Svensson* (1996), *Sargent* (1996), *Lucas* (*1996*) *and Chari* (1998).

Selected Works

1962. (With Z. Griliches, G.S. Maddala, and N. Wallace). Notes on estimated aggregate quarterly consumption functions. *Econometrica* 30, 491–500.

1967a. Optimal investment policy and the flexible accelerator. *International Economic Review* 8, 78–85.

1967b. Tests of a capital-theoretic model of technological change. *Review of Economic Studies* 34, 175–189.

1967c. Adjustment costs and the theory of supply. *Journal of Political Economy* 75, 321–334.

1968. (With T. McGuire, J. Farley and W. Ring.) Estimation and inference for linear models in which subsets of the dependent variable are constrained. *Journal of the American Statistical Association* 63, 1201–1213.

1969a. (With L. Rapping.) Real wages, employment, and inflation. *Journal of Political Economy* 77, 721–754.

1969b. (With L. Rapping.) Price expectations and the Phillips curve. *American Economic Review* 59, 342–350.

1970a. Capacity, overtime and empirical production functions. *American Economic Review* 60, 23–27.

1970b. (With L.A. Rapping et al.). Real wages, employment and inflation. In *The new microeconomics in employment and inflation theory*, ed. E.S. Phelps et al. New York: W.W. Norton.

1971. (With E.C. Prescott.) Investment under uncertainty. *Econometrica* 39, 659–681.

1972a. (With E.C. Prescott.) A note on price systems in infinite dimensional space. *International Economic Review* 13, 416–422.

1972b. Expectations and the neutrality of money. *Journal of Economic Theory* 4, 103–124.

1972c. (With L. Rapping.) Unemployment in the great depression: Is there a full explanation? *Journal of Political Economy* 80, 186–191.
1972d. Econometric testing of the natural rate hypothesis. In *The econometrics of price determination conference*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.
1973. Some international evidence on output-inflation trade-offs. *American Economic Review* 63, 326–334.
1974. (With E.C. Prescott.) Equilibrium search and unemployment. *Journal of Economic Theory* 7, 188–209.
1975. An equilibrium model of the business cycle. *Journal of Political Economy* 83, 1113–1144.
1976. Econometric policy evaluation: A critique. *Carnegie-Rochester conference series on public policy*, vol. 1. First publ. in *The Phillips Curve and Labor Markets*, ed. K. Brunner, and A. Meltzer. Amsterdam: North-Holland, 1975.
1977. Understanding business cycles. In *Stabilization of the Domestic and International Economy*, ed. K. Brunner, and A. Meltzer. Amsterdam: North-Holland.
1978a. Asset prices in an exchange economy. *Econometrica* 46, 1429–1445.
1978b. On the size distribution of business firms. *Bell Journal of Economics* 508–523.
1978c. Unemployment policy. *American Economic Review* 68, 353–357.
1979. (With T.J. Sargent.) After Keynesian macroeconometrics. In *After the Phillips curve*, Federal Reserve Bank of Boston, Conference series no. 19: 49–72. Repr. in *Federal Reserve Bank of Minneapolis Quarterly Review* 3, 1–6.
1980a. Rules, discretion and the role of the economic advisor. In *Rational expectations and economic policy*, ed. S. Fischer. Chicago: University of Chicago Press for the NBER.
1980b. Equilibrium in a pure currency economy. In *Models of monetary economics*, ed. J.-H. Karaken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.
1980c. Two illustrations of the quantity theory of money. *American Economic Review* 1970, 1005–1014.
1980d. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12, 696–717.
1981a. (With T.J. Sargent.) *Rational expectations and econometric practice*. Minneapolis: University of Minnesota Press.
1981b. *Studies in business-cycle theory.* Cambridge, MA: MIT Press.
1981c. Distributed lags and optimal investment policy. In Lucas and Sargent (1981a). 1981d. Optimal investment with rational expectations. In Lucas and Sargent (1981a).
1982. Interest rates and currency prices in a two-country world. *Journal of Monetary Economics* 10, 335–360.
1983. (With N.L. Stokey.) Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12, 55–94.
1984a. (With N.L. Stokey.) Optimal growth with many consumers. *Journal of Economic Theory* 32, 139–171.
1984b. Money in a theory of finance. In *Essays on Macroeconomic Implications of Financial and Labor Markets and Political Processes*, ed. K. Brunner, and A. Meltzer. Amsterdam: North-Holland.
1986a. Principles of fiscal and monetary policy. *Journal of Monetary Economics* 17, 117–134.
1986b. Adaptive behavior and economic theory. *Journal of Business* 59, S401–S426.
1987. (With N.L. Stokey.) Money and interest in a cash-in-advance economy. *Econometrica* 55, 491–514.
1987. *Models of Business Cycles*. Y. Jahnsson Lectures. Oxford: Basil Blackwell.
1988. On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.
1989. (With N.L. Stokey and E.C. Prescott.) *Recursive methods in economic dynamics*. Cambridge, MA: Harvard University Press.
1990a. Liquidity and interest rates. *Journal of Economic Theory* 50, 237–264.
1990b. Why doesn't capital flow from rich to poor countries? *American Economic Review* 80, 92–96.
1990c. Supply side economics: An analytical review. *Oxford Economic Papers* 42, 293–316.

1992a. (With A.G. Atkeson.) On efficient distribution with private information.
Review of Economic Studies 59, 427–453.
1992b. On efficiency and distribution. *Economic Journal* 102, 233–247.
1993. Making a miracle. *Econometrica* 61, 251–272.
1995. (With A.G. Atkeson.) Efficiency and equality in a simple model of efficient unemployment insurance. *Journal of Economic Theory* 66, 64–88.
1996. Nobel lecture: Monetary neutrality. *Journal of Political Economy* 104, 661–82.
2000a. Inflation and welfare. *Econometrica* 68, 247–274.
2000b. Some macroeconomics for the 21st century. *Journal of Economic Perspectives* 14(1), 159–168.
2001a. *Lectures on Economic Growth*. Cambridge, MA: Harvard University Press.
2001b. Externalities and cities. *Review of Economic Dynamics* 4, 245–274.
2002. (With E. Rossi-Hansberg.) On the internal structure of cities. *Econometrica* 70, 1445–1476.
2003. Macroeconomic priorities. *American Economic Review* 93, 1–14.
2004. Life earnings and rural-urban migration. *Journal of Political Economy* 112 (S1), S29–S59.
2007. (With M. Golosov.) Menu costs and Phillips curves. *Journal of Political Economy* 115, 171–199.
2007. (With F. Alvarez.) General equilibrium analysis of the Eaton–Kortum model of international trade. *Journal of Monetary Economics*.

Bibliography

Barro, R., and X. Sala-i-Martin. 2004. *Economic growth*, 2nd ed. Cambridge, MA: MIT Press.
Chari, V.V. 1998. Nobel Laureate Robert E. Lucas, Jr.: Architect of modern macroeconomics. *Journal of Economic Perspectives* 12(1), 171–186 Repr. in *Federal Reserve Bank of Minneapolis Quarterly Review* 23, 2–12, 1999.
Fisher, S. 1996. Robert Lucas's Nobel memorial prize. *Scandinavian Journal of Economics* 98: 11–31.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–15.
Hall, R.E. 1996. Robert Lucas, recipient of the 1995 Nobel memorial prize in economics. *Scandinavian Journal of Economics* 98: 33–48.
Kareken, J.A., T. Muench, and N. Wallace. 1973. Optimal open market strategy: The use of information variables. *American Economic Review* 63: 156–172.
Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.
Phelps, E. 1968. Money-wage dynamics and labor market equilibrium. *Journal of Political Economy* 76: 687–711.
Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.
Romer, P. 1986. Increasing returns and long run growth. *Journal of Political Economy* 94: 1002–1037.
Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the contrivance of money. *Journal of Political Economy* 66: 467–482.
Sargent, T.J. 1996. Expectations and nonneutrality of Lucas. *Journal of Monetary Economics* 37: 535–548.
Svensson, L.E.O. 1996. The scientific contributions of Robert E. Lucas, Jr. *Scandinavian Journal of Economics* 98: 1–10.

Lump Sum Taxes

J. de V. Graaff

Keywords

Excess burden of taxation; Initial endowments; Lump sum taxes; Poll tax; Redistribution of income of wealth

JEL Classifications

O1

A lump sum tax is fixed in amount and of such a nature that no action by the victim (short of emigration or suicide) can alter his or her liability. An example would be a poll tax, perhaps differentiated on the basis of sex and age.

It is difficult to find other examples. Differentiation on the basis of ability, wealth, income or expenditure would clearly lead to taxes that were not lump sum. Ability can be disguised. Wealth can be consumed. Leisure can be substituted for income, and saving for spending. All such actions would reduce tax. This implies that the principal criteria one might like to use as a basis for

redistributive taxation are ruled out if one is confined to lump sum taxes. It also implies that it may be difficult to relate lump sum taxes to ability to pay. A feature of lump sum taxation is that what taxpayers bear is exactly balanced (in monetary terms) by what the fisc gains. That is because there is no tax at the margin. (If there were tax at the margin, taxpayers could vary their liabilities by varying their activities, and the tax would not be lump sum.) The absence of tax at the margin means that no transaction is killed off by the driving of a wedge between what one party pays and the other receives.

When there is such a wedge (caused by a tax that is not lump sum) transactions are not entered into which, but for the tax, would have been mutually advantageous to the parties; and the loss to the parties is not balanced by any gain to the fisc. This is the 'excess burden' of taxation. It can never occur when taxes are lump sum.

In general equilibrium analysis the imposition of a set of lump sum taxes and bounties is equivalent to an adjustment of initial endowments. The attainment of equilibrium is not impaired, but its position will usually be altered. In welfare economics the conditions are investigated under which such an equilibrium may also represent a general optimum of production and exchange (in the sense of Pareto). If these conditions are met, it will not be possible to make one person better off without making someone else worse off. But the distribution of wealth may be very unequal: in an extreme case one person could end up with everything, the others with nothing. Lump sum taxes can, in theory, correct this situation without impairing the general optimum. In this sense, they are an ideal form of taxation.

Lump sum taxes are thus of some importance in theoretical work. But in the real world, poll taxes being their only viable form, they are rarely encountered precisely because they cannot in practice be matched to ability to pay or used to achieve a redistribution of income of wealth without ceasing to be lump sum. At most they are a benchmark against which the less than perfect taxes we normally encounter can be measured.

See Also

- ▶ Compensation Principle
- ▶ Neutral Taxation
- ▶ Optimal Taxation

Lundberg, Erik Filip (1907–1987)

Assar Lindbeck

Keywords

Cost inflation; Horndal effect; Inventory cycles; Inventory theory; Lundberg lag; Lundberg wage-multiplier; Lundberg, E. F.; Metzler, L. A.; Redistribution of income; Stabilization policy

JEL Classifications

B31

Lundberg was born in Stockholm and obtained a Ph.D. in economics in 1937 at the Stockholms Högskola. From 1937 to 1955 he was director of the Government Economic Research Institute (Konjunkturinstitutet), and from 1946 to 1965 he was professor of economics at the University of Stockholm; he held the same post at the Stockholm School of Economics from 1965 to 1970. He was president of the Royal Swedish Academy of Science from 1973 to 1976, and chairman of the Nobel Prize Committee for Economics from 1975 to 1980. He held numerous visiting professorships throughout the world.

Lundberg's main contributions to economic theory are his models of macroeconomic fluctuations and his analysis of the problems of economic policy, in particular the conflicts between stabilization policy and policies for the allocation of resources and the distribution of income.

His *Studies in The Theory of Economic Expansion* (1937) is an early work of high originality about the instability of growth, the main analytical technique being systems of difference equations

of multiplier and accelerator mechanisms (with some consideration to the possibilities of flexible coefficients), embedded in a simple macroeconomic framework. Lags between inputs, output, income formation and spending play strategic roles (the lag between output and income-formation is often referred to as 'the Lundberg lag'). Rather than providing reduced-form solutions to the system, Lundberg presented numerical sequences of various macroeconomic variables and their relations, so-called 'sequence analysis'.

The part of the book which had the strongest immediate influence on other theorists is perhaps the inventory model. Non-anticipated increases in sales, while first resulting in a fall in inventories, later on, due to attempts by firms to restore the initial relation of inventory stocks to production levels, result in various kinds of inventory cycles. Lundberg's inventory analysis inspired, for instance, Lloyd Metzler's inventory model, as well as the inventory analysis, with more elaborate microeconomic underpinnings, by Holt and Modigliani.

Among Lundberg's contributions to the analysis of economic policy, *Business Cycles and Economic Policy* (1953 in Swedish; 1957 in English), stands out as a particularly important piece of work. The analysis is characterized by rather informal theorizing, though using concepts of traditional economic theory, both for the 'international' and the Swedish economy. Calculations of *ex ante* inflation and deflation gaps, by way of excess demand (supply) in the goods market and/or the labour market, are important instruments of analysis.

Lundberg was also a pioneer in analysing the role of taxation for 'cost inflation', a point formalized by an equation expressing how much nominal wage rates would have to rise to guarantee a one per cent increase in after-tax real wage rates, after considering both the marginal tax rates and the price effects of wage increases (the Lundberg wage-multiplier).

When analysing long-term growth problems, Lundberg also discovered the so- called 'Horndal effect', expressing how labour productivity can go on rising over long periods of time without new investment, hence providing an indication of disembodied productivity growth (1961). Lundberg also made interesting comparative studies of growth, fluctuations and economic policy in various countries, for instance in *Instability and Economic Growth* (1968).

He also participated frequently in Swedish economic policy discussion, emphasizing the importance of avoiding overvalued exchange rates. His own policy recommendations, in addition, built on combining *general,* market-orientated stabilization policies with rather selective social policies to achieve economic security and desired income redistributions.

Selected Works

1930. Om ekonomisk jämvikt. *Ekonomisk Tidskrift* 32(4): 133–160.
1937. *Studies in the theory of economic expansion.* London: P.S. King & Sons. Oxford: Blackwell, 1955.
1953. *Konjunkturer och Ekonomisk Politik.* Stockholm: SNS. English edn, *Business cycles and economic policy.* London: Allen & Unwin, 1957.
1959. The profitability of investment. *Economic Journal* 69: 653–677.
1961. *Produktivitet och räntabilitet.* Stockholm: SNS.
1968. *Instability and economic growth.* New Haven: Yale University Press.
1972. Productivity and structural change – A policy issue in Sweden. *Economic Journal* 82(Supplement): 465–485.
1985. The rise and fall of the Swedish model. *Journal of Economic Literature* 23: 1–36.

Lutz, Friedrich August (1901–1975)

Jürg Niehans

Lutz was born on 29 December 1901, in Sarrebourg (Lorraine), then part of Germany. He died in Zurich on 4 October 1975. After studying economics in Heidelberg, Berlin and

Tübingen and working briefly for the German Machine Builders' Association, he embarked on an academic career as an assistant to Walter Eucken in Freiburg (Germany), where he was a lecturer from 1932 to 1938. He belonged to the neoliberal 'Freiburg Circle' for the rest of his life and later became one of the founders of the Mont Pélerin Society.

Finding academic life in Hitler's Germany incompatible with his liberal views, Lutz emigrated to the United States in 1938. Starting again as an instructor in Princeton, he rose through the ranks to become a professor in 1947. This was scientifically his most productive period. Important contributions were made jointly with his wife, Vera Smith Lutz, a respected economist in her own right. From 1953 until his retirement in 1972, Lutz was Professor of Economics at the University of Zurich. Tübingen gave him an honorary degree in 1967.

Lutz was chiefly concerned with money, capital and interest. His term-structure paper (1940) provided a synthesis of the 'neoclassical' view that long rates, in equilibrium, are an average of short rates, but his formula differs from that of Hicks inasmuch as it determines the coupon that makes the present value of a long-term bond equal to its par value.

Lutz's *Theory of Investment of the Firm* (1951) extends the microeconomic theory of the firm to time and capital. It is representative for the level of analysis of the 1950s in the same way that Hirshleifer's *Investment, Interest and Capital* (1970) is representative for the Seventies and Hayek's *Pure Theory of Capital* (1941) for the Thirties. While the taxonomy of innumerable 'cases' makes difficult reading, the maximization of the present value of quasi–rents clearly emerges as the unifying principle. In his *Zinstheorie* (1956) (*The Theory of Interest*, 1967), Lutz provides an authoritative and critical survey of interest theory since Böhm–Bawerk.

In later life, Lutz was a highly respected expert on international monetary policy and, like so many non-Keynesians of the Keynesian generation, an ardent advocate of floating exchange rates. A gentle man, he combined firmness in his convictions with respect for those of others.

Selected Works

1932. *Das Konjunkturproblem in der Nationalökonomie*. Probleme der theoretischen Nationalökonomie, vol. 2, ed. by W. Eucken. Jena: Gustav Fischer.

1938. The outcome of the saving–investment discussion. *Quarterly Journal of Economics* 52. Reprinted in American Economic Association, *Readings in business cycle theory*. Philadelphia and Toronto: Blakiston, 1944.

1940. The structure of interest rates. *Quarterly Journal of Economics* 55. Reprinted in American Economic Association, *Readings in the theory of income distribution*. Philadelphia and Toronto: Blakiston, 1946.

1945. *Corporate cash balances 1914–1943, manufacturing and trade*. New York: National Bureau of Economic Research.

1947. (With N.S. Buchanan) *Rebuilding the world economy; America's role in foreign trade and investment*. New York: Twentieth Century Fund.

1951. (With V.S. Lutz) *The theory of investment of the firm*. Princeton: Princeton University Press. Reprinted, New York: Greenwood Press, 1969.

1956. *Zinstheorie*. Hand– und Lehrbücher aus dem Gebiet der Sozialwissenschaften, ed. E. Salin and G. Schmölders. Zurich/Tübingen: Polygraphischer Verlag/Mohr (Siebeck). 2nd revised ed, 1967. Japanese trans. 1962; English trans. as *The theory of interest*. Dordrecht: D. Reidel, 1967.

1961. The essentials of capital theory. In *The theory of capital, proceedings of a conference held by the international economic association*, ed. F.A. Lutz and D.C. Hague. London: Macmillan, 1961.

1962. *Geld und Währung: Gesammelte Abhandlungen*. Walter Eucken Institut, Wirtschaftswissenschaftliche und wirtschaftsrechtliche Untersuchungen, vol. 1. Tübingen: Mohr (Siebeck).

1971. *Politische Ueberzeugungen und nationalökonomische Theorie: Zürcher Vorträge*. Walter Eucken Institut, Wirtschaftswissenschaftliche und wirtschaftsrechtliche Untersuchungen, vol. 8. Tübingen: Mohr (Siebeck). With list of publications.

Luxemburg, Rosa (1870–1919)

Tadeusz Kowalik

Rosa Luxemburg was born on 5 March 1870 in Zamosc (Polish territory under the Russian occupation), and died, murdered during the revolution, on 15 January 1919 in Berlin. Rosa Luxemburg was a socialist thinker and writer, one of the leaders of Polish and German Social Democracy, and an economist. She studied in Zurich, first philosophy and natural sciences (for two years), then she graduated from the Faculty of Law and Economics. In 1897 she received her PhD for a book *Die industrielle Entwicklung Polens* (The Industrial Development of Poland) (1898). In 1898 she contracted a marriage of convenience (with G. Luebeck) to obtain German citizenship and from then, until the end of her life, she lived in Berlin. She was one of the founders of the Social Democratic Party in the Kingdom of Poland (under the Russian occupation). The main area of her activity was German Social Democracy, in which she became one of the leading intellectuals. Her articles in which she opposed the revisionism of Eduard Bernstein and defended revolutionary Marxism won her European popularity. (They were subsequently published in a book *Sozialreform oder Revolution?* [1900].)

In the period 1907–14 Luxemburg lectured in political economy, then in economic history in the Party School of German Social Democracy (her predecessor lecturing in political economy was Hilferding). Towards the end of that period Luxemburg elaborated her lecture notes to publish them in the form of a manual, but in view of the theoretical problems she encountered, she left the manuscript unfinished. Half the chapters were lost during the war, and the remainder were published under the title *Einführung in die Nationaloekonomie* (1925).

She aimed at producing an orthodox, popularizing manual. In the process of doing this she was still convinced that political economy found its 'peak and climax' in Marx's works and that it could be developed by his followers 'only in details'. Attempting to give an outline of the general tendencies of the capitalist economy however, she faced insurmountable problems, previously unsuspected. She could find no satisfactory answer in Marx to the question 'what are the objective historical limits to capitalism?' Excited by her own hypothesis, she wrote 'wie im Rausch': in a period of four months she produced over five hundred pages and 'without even once reading the draft' turned it over to the publisher. This was the genesis of her *opus magnum – Die Akkumulation des Kapitals* (1913; English trans., 1951).

One could make a figurative comparison of her four-month effort to the activity of a volcano ejecting a flow of ideas, with its trains of thought picked up and abandoned, its questions left with no answer, its contradictory contentions. Hence there are tremendous problems in interpreting the results.

One of several possible interpretations is as follows: The evolution of her ideas, and particularly the 'illumination of 1912' exemplifies a more general trend in the development of Marxian economic thought after Marx's death. One can distinguish in the trend two rather different currents. The first is based on Marxian theory of value and surplus-value. This current has been developed mainly by the first generation of Marxists, such as Karl Kautsky. The second current came to the fore rather later. The most representative figures are Hilferding and Rosa Luxemburg in Germany, and Lenin and Tugan-Baranovsky in Russia. They undertook the task of developing those aspects of Marx's theory that deal with the dynamics of modern capitalist economy. The year 1912 marks the border between Rosa Luxemburg the 'orthodox' and Rosa Luxemburg the 'revisionist', if we use this label in a theoretical rather than a political sense. Rosa Luxemburg's changed attitude toward the Marxian theory of capitalism manifests itself in a change in her methodology. In the *Einführung* she used the method that Marx applied in the first volume of *Das Kapital,* where the starting point was an analysis of the individual commodity and individual capital. The very essence of the turning-point in her later economic

theorizing consists in grasping the importance of a macroeconomic approach. She became fascinated by Marx's concept of global reproduction and accumulation, developed in the second volume of *Das Kapital* (Marx's schemes of reproduction). In this construction she now saw the perfect embodiment of Marxist political economy and the most powerful analytical tool. Francois Quesnay was now advanced, in her eyes, to the rank of a founder of economics as an exact science, while she blamed the English line of classical economy for completely obscuring the eternal and universal functions of the means of production in the labour process.

In Rosa Luxemburg's thinking, fascination with the Marxian schemes of reproduction as a promising tool of analysis of a capitalist system as a whole goes together with the argument that the decisive part of *Das Kapital* (the last part of the second volume) was unfinished and underdeveloped. In the form left by Marx, and published posthumously by Engels, the model of accumulation had been constructed, in her opinion, on several drastic assumptions which make it impossible to understand the nature of capitalist development and of its limits.

The model assumes an identity of production and realization, which means that capitalist production creates a sufficiently large sales market for itself. This assumption contradicts not only the spirit of Marx's theory but also many statements in the first and third volumes of *Das Kapital* about a tendency on the part of total demand to lag behind rapidly increasing production.

Moreover, this assumption is related to the next great disadvantage of Marx's scheme: disregard of the circulation of money. As a consequence of this, Marx could not draw any satisfactory analytical conclusions from his rejection (in the first volume of *Das Kapital)* of Say's Law. In modern terms we could say that in disregarding money Marx identified savings with real accumulation (investment).

Marx analysed the accumulation of capital within a framework of society composed only of the capitalist and the working class. In Luxemburg's opinion, this assumption of pure capitalism rendered impossible the discovery of which class benefits from the expansion of capitalist production. Approached from this angle the Marxian model of reproduction can only be understood as a vision of production for production's sake.

Another disadvantage of the Marxian concept is the assumption of unchanged organic composition of capital and constant productivity of labour. As was the case with many Marxists of her day, Luxemburg recognized only one type of technical progress, what is now called 'capital-intensive'. She was convinced that technical progress must manifest itself in an increasing organic composition of capital, i.e. increasing share of constant capital in the value of the product, or, what was for her only another way of expressing the same phenomenon, in an increasing share of Division I (the production of the means of production) in the total social product.

Luxemburg promised much more than she was able to deliver. At different stages of her analysis she tried to overcome Marx's shortcomings. However, she did not succeed in transforming the schemes of reproduction into a form which would suit her purpose. For example, the criticism of Marx's concept of pure capitalism runs through the whole of her book, but whenever she resorts to the schemes of reproduction she uses them in the original (Marx's) form. The only correction made by her to the Marxian construction was that she allowed for an increase in the productivity of labour: in her schemata of reproduction the organic composition of capital *(c/v)* increases from period to period. On this ground she argued that expanded reproduction inevitably brings an increasing deficit of the means of production and an increasing surplus of the means of consumption. Disproportions arising because of that could be, according to her conviction, liquidated or dampened only outside the framework of pure capitalism – by exchange between capitalist and pre-capitalist systems.

This conclusion was based not only on her general law of increase of organic composition of capital, but also on the erroneous conviction that accumulation must be allocated to the division in which it has been obtained. Thus, in her only attempt to introduce corrections in the

Marxian schemata of expanded reproduction, Luxemburg cannot claim any visible theoretical achievements in the analysis of capitalist accumulation. No conclusions resulted from this attempt concerning her main contention: lack of sufficient demand as the crucial obstacle of capitalist development. In this sense the book is disappointing.

However, her work cannot be neglected. The significance of *The Accumulation of Capital* lies in the fact that it is an attempt at a theoretical solution of the known Marxian statement that the conditions of production are not identical with the conditions of realization. By rejecting Say's law she tried to prove that accumulation is affected to a large extent by the prospect of a growing market which, in turn, is determined primarily by the existing sales situation. Thus, pure capitalism provides by itself too weak a basis for rapid economic growth. Saving does not transform itself automatically into real investment. This was the direction of development of a theory of capitalist dynamics in the following decades. Michal Kalecki (1971) was the most successful in taking up problems posed by Rosa Luxemburg and solving them effectively. But, due to some special historical conditions, Marxists for a long time treated Kalecki's theory with suspicion or indifference.

An attempt by Luxemburg to include the monetary system in the theory of capitalist reproduction and accumulation also deserves attention. It can be seen from numerous passages in the second volume of *Das Kapital* that Marx tried, but failed, to solve this tremendously difficult question. It is true that Rosa Luxemburg did not solve it either. But in contrast to many other disciples of Marx she did not neglect the problem and formulated it in a much more lucid and precise way than all her predecessors and contemporaries.

Why did Rosa Luxemburg raise again the problem of the incentives to accumulation, investment and technical progress in the capitalist system? What led her to the conviction that Marx's analysis is not sufficient? One can suppose that there were historical reasons, as well as theoretical issues. Her discussion with Eduard Bernstein at the turn of this century may provide one possible explanation. She then expressed the following view:

> In the general development of capitalism small capital ... plays the part of the pioneer of technical revolution. ... If small capital is the champion of technical progress and if technical progress is the pulse of a capitalist economy then small capital is a phenomenon inseparable from capitalist development ... The gradual disappearance of medium-sized firms would not mean, as Bernstein seems to think, that the development of capitalism is revolutionary, but on the contrary, it would indicate, that it is stagnant and drowsy *(Social Reform or Revolution?,* part I:2 Adaptation of Capitalism).

Some dozen years later it was clear to her that the capitalist economy was entering the era of industrial giants and that the individual entrepreneurs of the period of free competition and the corresponding mechanisms were beginning to fade away. Rosa Luxemburg must have asked herself: 'Why does capitalism not show signs of stagnation despite this process of structural transformation?' The explanation given by Marx that the capitalist strives incessantly to maximize his profits and in the conditions of free competition this striving becomes for each individual capitalist the 'external law of compulsion' was not valid for the new conditions.

We already know the general tenor of her theoretical answer: neither the consumption fund of the working class nor the consumption expenditures of the capitalists can provide sufficiently strong incentives to accumulation. A large part of the incentive to accumulation in a capitalist system is due to a steady and uninterrupted economic exchange between capitalist and non-capitalist environments.

Historical studies led Rosa Luxemburg to the conviction that there was no 'Chinese Wall' between classical capitalism and the phase of imperialism. This was so because political violence was for her 'nothing but a vehicle for the economic process' (1951, p. 452). Seeing ever more clearly the important of non-economic factors for capitalist accumulation in the past and in the future, she advanced to a broader interpretation of the process of the development of capitalism than the interpretation given by Marx in *Das Kapital.* Capital is not only born 'soaked in blood and dirt' (Marx), but grows later in very much the same way, until the moment of its collapse. Thus

Rosa Luxemburg's interpretation of the essence and character of imperialism is very broad. First, a period of wars and revolutions arising from the exhaustion of the non-capitalist system provides external markets for capitalist accumulation, areas for the profitable investment of capital, and basic raw materials. Without this environment as a 'feeding ground', accumulation would be, in her opinion, impossible. The main achievement of *Accumulation of Capital* probably lies in locating the problem of underdeveloped countries as a central issue in the debate on the further development or collapse of the capitalist system.

The title of the last chapter of *The Accumulation of Capital* is 'Militarism as a sphere of the accumulation of capital'. Rosa Luxemburg makes an attempt to analyse the importance of armaments production – as production and not as a tool of external expansion – for stimulating economic growth in capitalism.

She rejected the conviction prevailing at that time that the bourgeois state can merely redistribute profits and incomes, without changing anything in the conditions of reproduction of total social capital. Government expenditures for armament production resulted in the state creating 'by sleight of hand' new demand, new purchasing power, and thus influencing the magnitude of the total accumulation of capital. The demand created in this way by the state has the same effect as a newly opened market. In the era of imperialism, armament production becomes one of the important ways of solving difficulties in the realization of growing production. The attractiveness of expanding this sphere of accumulation consists, in addition, in the fact that expenditure by the state in military equipment 'free of the vagaries and subjective fluctuations of personal consumption, it achieves almost automatic regularity and rhythmic growth' (1951, p. 466). Moreover, military expenditure places a lever with automatic and thythmic movement in the hands of capitalist state, so that it seems at first capable of infinite expansion.

Needless to say, from today's point of view Rosa Luxemburg's approach with regard to the general role of the state is rather narrow. She also did not perceive the possibility of credit creation by budget deficits. The multiplier effect of the armament sector was hardly noticed. It is not clear whether she assumed unused productive capacity. Too much stress was laid on wages and individual incomes of small producers, as main sources of government revenue.

But the mere fact of raising the problem, considered very important today, and of showing the fundamentally correct direction in which its solution should go, elevates her to the rank of precursors of contemporary economic thinking.

The question of the collapse of the capitalist system plays an important part in Rosa Luxemburg's thinking. The desire to grasp theoretically the objective historical limits of the mode of production was one of her motives in dealing with the problem of accumulation. In *The Accumulation of Capital,* and in her *Anti-Critique* she often returns to this problem. As an historical process the accumulation of capital, according to her, depends 'in every respect upon the non-capitalist social strata and forms of social organization' (1951, p. 366). In this way, the solution of the problem that had been a subject of controversy since the time of Sismondi, according to whom the accumulation of capital is altogether impossible, and the naive optimism of Say and Tugan-Baranovsky, in whose opinion capitalism can fertilize itself *ad infinitum,* is in dialectical contradiction which is expressed in the fact that the environment of non-capitalistic social formations is essential for the accumulation of capital and that only by the exchange with them can it progress and last as long as this environment exists.

This last thought, and her contention that accumulation internationalizes the capitalist mode of production by eliminating the traditional modes of production and, at the same time, cannot survive in pure capitalism, is repeated several times. However, this is only an abstract point, not a comprehensive concept of the breakdown of the capitalist system; only a 'theoretical formulation' showing a tendency in the development of capitalism – and nothing else. She made her abstract thesis on the impossibility of the existence of capitalism without the pre-capitalist environment more specific by her historical analysis of economic and socio-

political conflicts of interests between the 'imperialist' and 'colonial' countries as a primary source of wars and revolutions.

Selected Works

1898. *Die Industrielle Entwicklung Polens*. Leipzig.
1900. *Sozialreform oder Revolution?*. Leipzig.
1913. *Die Akkumulation des Kapital. Ein Beitrag zug ökonomischen Erklärung des Imperialismus*. Berlin. Trans. by Agnes Schwarzschild as The Accumulation of Capital. London: Routledge & Kegan Paul, 1951 (with a foreword by Joan Robinson).
1921. *Die Akkumulationm des Kapitals, oder was die Epigonen aus der Marxschen Theorie gemacht haben*. Berlin: Eine Antikritik.

References

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy 1933–1970*. Cambridge: Cambridge University Press.
Kowalik, T. 1964. Rosa Luxemburg's theory of accumulation and imperialism. In *Problems of economic dynamics and planning, essays in honour of Michal Kalecki.* Warsaw: Panstowowe Wydawnictwo Naukowe, 1964. (A summary of a work available in Polish, Italian and Spanish.)
Nettl, J.P. 1966. *Rosa Luxemburg*, vol. 2. Oxford: Oxford University Press.

Lyapunov Functions

C. Henry

JEL Classifications
C0

Within twelve years, from Poincaré's *Mémoire sur les courbes définies par une équation différentielle* (1881–1886) to Lyapunov's thesis *Obshčaya zadača ob unstoičivosti dviženiya* (1892), the qualitative theory of differential equations emerged almost from scratch as the core of a new field in mathematics; both Poincaré and Lyapunov were motivated by problems in mechanics, celestial mechanics above all. Even if he did not match Poincaré's prodigious creativity between 1880 and 1883, Lyapunov developed from 1888 to 1892 a theory of dynamical stability which makes his 1892 thesis both a pioneering piece of work and a classic; in particular he developed a general stability criterion which now bears his name: the Lyapunov function.

Consider a system of ordinary differential equations

$$\dot{x} = f(x,t)$$

Where x is a vector in R^n and depends on t (t is in general interpreted as time), where $\dot{x} = dx/dt$ is the derivative of x with respect to t and where f is a function from R^{n+1} to R^n. A trajectory of the system is a function x from an interval T in R to R^n

$$x : T \to R^n : t \to x(t)$$

$$\dot{x}(t) \equiv f(x(t),t)$$

which is a solution of the system, i.e. such that $\dot{x}(t)$ and $f[x(t), t]$ are identical on T; T is often of the form $[t_0, +\infty]$.

In what follows we shall limit ourselves to autonomous systems, i.e. systems of the form $\dot{x} = f(x)$, where f is dependent on t only through x. However, our whole presentation is easily generalized to non-autonomous systems, as is done in Rouche et al. (1977) and Rouche and Mawhin (1980).

It would appear at first sight that the system $\dot{x} = f(x)$ suffers from another restriction: it is a first-order system, in the sense that its equations include first-order derivatives only. This might be seen a serious restriction indeed; think for example of the system formalizing the dynamics of the simple frictionless pendulum

$$\ddot{x}_1 + \sin x_1 = 0$$

where x_1 is the angular distance from the vertical line. However, it is always possible to transform a

system including derivatives of order higher than one into a first-order system with a higher number of equations. For example, the former system consisting of one second-order equation is equivalent to the following system of two first-order equations:

$$\dot{x}_1 = x_2 \dot{x}_2 = -\sin x_1$$

To investigate the stability properties of the pendulum, it is thus immediately possible to make use of the general concepts and methods available for first-order systems.

In order to introduce these concepts and methods in the spirit of Lyapunov, and then to see how they operate in economic models, we first have to be slightly more precise in defining an autonomous differential system as a system of first-order differential equations

$$(\mathrm{DS}) \qquad \dot{x} = f(x)$$

where f is a continuous Lipschitzian function from an open subset Ω of R^n to R^n, i.e.

$$f : \Omega \to R^n : x \to f(x).$$

'Lipschitzian' means that there exists a constant α such that

$$\forall x^1, x^2 \in \Omega, \left\| f\left(x^1\right) - f\left(x^2\right) \right\| \leq \alpha \left\| x^1 - x^2 \right\|;$$

this assumption is very convenient because it ensures (see Coddington and Levinson 1955) that through any point in Ω there passes one and only one trajectory of (DS); hence trajectories do not cross. However this assumption is not strictly necessary for what follows (see Aubin and Cellina 1984; Rouche et al. 1977).

We may now introduce the basic concepts of stability and attractivity for equilibria of dynamic systems. A point x^e in Ω is an equilibrium of (DS) if $f(x^e) = 0$; in other words, x^e is an equilibrium if for any t_0 in R the function

$$x : [t_0, +\infty[\to R^n : t \to x^e$$

is a trajectory.

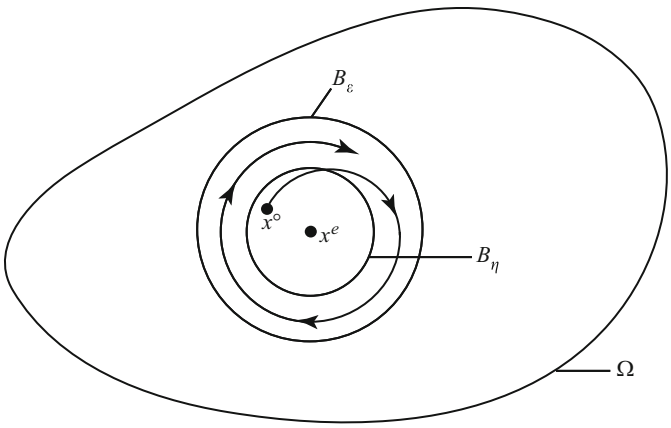
Stability: an equilibrium x^e is stable if a trajectory which comes sufficiently close to x^e never after recedes too far from x^e.

More precisely an equilibrium x^e is stable if, for any neighbourhood B_δ of x^e included in Ω, there exists a neighbourhood $B_\eta \subset B_\varepsilon$ such that any trajectory passing through B_η remains in B_ε ever after (see Fig. 1).

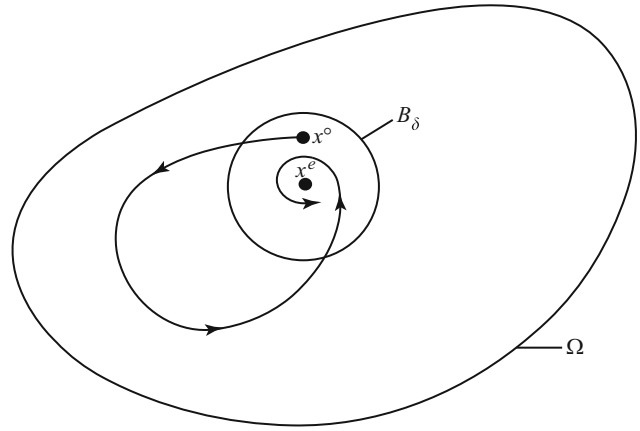
Local attractivity: an equilibrium x^e is a local attractor if a trajectory which comes sufficiently close to x^e later on tends to x^e.

More precisely an equilibrium x^e is a local attractor if there exists a neighbourhood B_δ of x^e included in Ω such that any trajectory which passes through B_δ tends to x^e as $t \to +\infty$; this

Lyapunov Functions, Fig. 1



Lyapunov Functions, Fig. 2



does not mean that the trajectory always remains in B_δ (see Fig. 2).

An equilibrium may be stable without being a local attractor, as in the case of the frictionless pendulum. It is also true that an equilibrium may be a local attractor, without being stable, but this is much more difficult to illustrate (see section 40 in Hahn 1967). An equilibrium which is both stable and a local attractor is often called asymptotically stable.

Global attractivity: given a subset Ω_s of Ω, an equilibrium x^e in Ω_s is a global attractor with respect to Ω_s if any trajectory which passes through Ω_s tends to x^e as $t \to +\infty$.

An equilibrium which is both stable and a global attractor with respect to some Ω_s is often called globally asymptotically stable (globally with respect to Ω_s).

The convenient way, often the sole way, to deal with stability and attractivity as defined above, is in general to find a suitable Lyapunov function.

Lyapunov function: consider a subset Δ of Ω and a function of class C^1 (i.e. continuous and having continuous first-order partial derivatives)

$$W : \Delta \to R : x \to W(x).$$

W is a Lyapunov function if it satisfies the following requirements:

(i) it is bounded below on Δ, i.e.

$$\exists a \epsilon R \text{ such that}, \forall x \epsilon \Delta, W(x) \geq a.$$

(ii) it tends to infinity as x does, i.e.

$$\text{if } \|x\| \to +\infty, \text{then } W(x) \to +\infty$$

(iii) its time derivative $\dot{W}(x)$ is nonpositive on Δ, i.e.

$$\forall x \epsilon \Delta, \quad \dot{W}(x) \leq 0$$

where the time derivative is defined as

$$\dot{W} : \Delta \to R : x \to \dot{W}(x) = \sum_{k-1}^{x} \frac{\partial W(x)}{\partial x_k} f_h(x).$$

The name 'time derivative' is warranted as

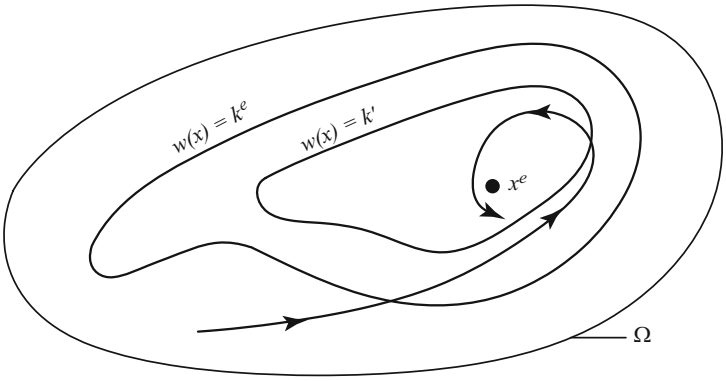
$$\dot{W}[x(t)] = \sum_{k-1}^{x} \frac{\partial W[x(t)]}{\partial x_k} \dot{x}_k(t) = \frac{dW[x(t)]}{dt}$$

along any trajectory in Δ.

It is possible for $n = 2$ to draw level curves of W in the subset Δ of the plane (x_1, x_2). On Fig. 3 two level curves are drawn, with $k' < k''$, as well as a trajectory.

W is non-increasing along any trajectory in Δ; hence, as soon as a trajectory reaches a level curve, say k', it never again comes back to points on level curves k with $k > k'$, for example $k = k''$.

Lyapunov Functions, Fig. 3



This suggests deriving properties of stability and attractivity from the existence of a Lyapunov function; we indeed have:

Proposition 1 if there exists on some neighbourhood B of the equilibrium x^e a Lyapunov function W and if x^e is an isolated minimum of W on B, then x^e is a stable equilibrium. If moreover x^e is the only point in B where $\dot{W} = 0$, then x^e is also a local attractor.

These are sufficient conditions for stability and local attractivity. It turns out that the existence of a Lyapunov function is also a necessary condition (for a general exposition and complete proofs which are valid even for nonautonomous systems, see Rouche et al. (1977) and Rouche and Mawhin (1980)).

The first part of Proposition 1, but not the second part, applies to the frictionless pendulum, the Lyapunov function being here the total energy $\frac{1}{2}x_2^2 - \cos x_1 + 1$. Both parts of Proposition 1 apply to the tâtonnement process in a competitive economy where all goods are gross substitutes for all prices. Let n be the number of goods; let p be the price vector normalized in such a way that it is in the $n - 1$ dimensional unit simplex $\overline{\Sigma}$ defined by $\sum_{j=1}^{n} Pj = 1$; and let $z_j(p)$ be the aggregate excess demand function for good j. Let Σ be the interior of $\overline{\Sigma}$. Gross substitutability implies that there exists one and only one general competitive equilibrium price vector p^e and that p^e is in Σ, i.e. p_j^e is strictly positive for all goods j (for more details see Arrow and Hahn 1971).

Consider then the well-known tâtonnement process

$$(\mathrm{TP})\dot{p} = z(p).$$

It is a (DS) system, with $\Omega = \Sigma$; being the unique general competitive equilibrium price vector, p^e is also the unique equilibrium of the differential system (TP).

Consider on $\sum$ the function

$$W(p) = ||p - p^e||^2.$$

Its time derivative is

$$\dot{W}(p) = \sum_{j-1}^{x} \frac{\partial W(p)}{\partial p_j} z_j(p)$$
$$= 2\sum_{j-1}^{x} \left(p_j - p_j^e\right) z_j(p) = -2p^e \cdot z(p),$$

because of Walras's law. On the other hand, it is a consequence of gross substitutability that

$$\forall p \neq p^e, p^e \cdot z(p) > 0.$$

It is thus clear that W is a Lyapunov function, that p^e is its unique minimum on Σ and that $\dot{W}$ is zero only at p^e; hence p^e is stable and is a local attractor for the tâtonnement process. Is it a global attractor with respect to Σ? The answer is not within the range of proposition 1. Something more is needed.

Proposition 2 consider a system (DS) and a bounded subset Δ of Ω which is such that there exists a Lyapunov function W

$$W : \Delta \to R : x \to W(x)$$

satisfying the additional requirement that $\dot{W}$ is zero only at equilibria of the system, i.e.

$$\dot{W}(x) = 0 \Rightarrow f(x) = 0.$$

Then, if all the limit points of a trajectory are in Δ, they are equilibria of the system; if moreover all the equilibria of the system are isolated, this trajectory tends to one of them as $t \to +\infty$.

Corollary if the system has a unique equilibrium, if Δ is open and if any trajectory which passes through Δ has all its limit points in Δ, then the equilibrium is a global attractor with respect to Δ and it is stable.

This corollary has no general counterpart when there are several isolated equilibria, because a trajectory starting in the neighbourhood of one equilibrium may tend to another one. However, if an equilibrium is an isolated minimum of W, prpt 1 ensures that it is stable and is a local attractor.

Proposition 2 and its corollary allow us to answer the qst, left unanswered above, about the tâtonnement process: as Σ is bounded and as gross substitutability prevents any trajectory from having a limit point on the boundary of Σ, all conditions in prpt 2 and in the corollary are met; so p^e is a global attractor with respect to Σ.

Another well-known application of Lyapunov functions is in the theory of public goods. Consider an economy with N consumers ($i = 1, \ldots, N$), m public goods ($k = 1, \ldots, m$) and one private good used as numeraire. Let $x \in R_+^m$ denote the bundle of public goods made available to the consumers, and let y^i denote the amount of numeraire consumed by $i = 1, \ldots, N$; this means that $(x, y^i) \in R_+^{m+1}$ describe the total consumption of i. His preferences are formalized by a utility function

$$u^i : R_+^{m+1} \to R : \left(x, y^i\right) \to u^i\left(x, y^i\right);$$

this function is of class C^1, quasi-concave, non-decreasing with respect to each of its arguments, and is strictly increasing with respect to the consumption of the numeraire, i.e.

$$\frac{\partial u^i}{\partial y^i} > 0 \text{ on } R_+^{m+1}.$$

Let Z be the set of feasible allocations, i.e. the set of all $z = (x, y^1, \ldots, y^N)$ in R^{m+N} which can be made available for consumption, given the technical possibilities and the initial resources of the economy. Z is of course bounded; it is reasonable to consider that it is closed and convex; hence it is a compact convex subset of R_+^{m+N}.

How to reach a Pareto-optimal feasible allocation? The MDP (for Malinvaud–Drèze–Poussin, see Champsaur et al. 1977) planning procedure gives the following answer: starting from any feasible allocation, revise z continuously according to the following differential system, which is a (DS) system:

$$\text{(MDP)} \quad \frac{\mathrm{d}x_k}{\mathrm{d}t} = \sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z), \; k = 1, \ldots, m$$

$$\frac{\mathrm{d}y^i}{\mathrm{d}t} = -\sum_{k=1}^{m} \pi_k^i(z) - \frac{\mathrm{d}x_k}{\mathrm{d}t} + \delta^i \sum_{k=1}^{m} \frac{\mathrm{d}x_k}{\mathrm{d}t} \left[\sum_{j=1}^{N} \pi_k^j(z) - \gamma_k(z) \right], \quad i = 1, \ldots, N$$

where $\pi_k^i(z)$ is the marginal willingness to pay of consumer i for public good k, and $\gamma_k(z)$ is the marginal cost of public good k. The δ_i, $i = 1, \ldots, N$, are non-negative weights summing up to $1 : \sum_{i=1}^{N} \delta^i = 1$. These differential equations mean that the quantity made available of each public good is revised according to the difference between the total marginal willingness to pay for that public good and its marginal cost; simultaneously every consumer pays an amount of numéraire equal to his willingness to pay for this set of revisions, and receives a fraction of the total surplus that the revisions generate.

Consider the function

$$W : Z \to R : z \to W(z) = -u^i\left(x, y^i\right)$$

where i is chosen among those i for which $\delta^i > 0$. Straightforward calculations lead to

$$\dot{W}(z) = -\delta^i \sum_{k-1}^{m} \left[\pi_k^i(z) - \gamma_K(z)\right]^2 \frac{\partial u^i}{\partial y^i},$$

which is nonpositive everywhere on Z and is zero if and only if z is an equilibrium of (MDP). W is a Lyapunov function and Proposition 2 applies with $\Delta = Z$. As Z is compact it is even possible to conclude that any limit point of any trajectory which is included in Z is an equilibrium of (MDP). If all the utility functions u^i, $i = 1, \ldots, N$, are strictly quasi-concave, the result is sharpened in the sense that any trajectory which is included in Z tends to an equilibrium as $t \to +\infty$. The economic significance of these results proceeds from the fact that all the equilibria of (MDP) are Pareto optima.

However, it is not guaranteed that all trajectories of (MDP) starting in Z are included in Z, as the revisions generated by the equations

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = \sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z), k = 1, \ldots, m$$

may lead to negative values of the public goods. We would then have a meaningless procedure. In order to avoid this possibility the above equations must be replaced in (MDP) by:

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = \begin{cases} \sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z) & \text{for } x_k > 0 \\ \max\left[\sum_{i=1}^{N} \pi_k^i(z) - \gamma_k(z), 0\right] & \text{for } x_k = 0. \end{cases}$$

It is then immediate that any trajectory of (MDP) starting in Z is included in Z. But do trajectories still exist? and if they exist, do they actually tend to equilibria of (MDP)? The answers are not trivial, as there are significant discontinuities in the right-hand sides of the new equations. These answers nevertheless turn out to be positive; this is a by-product of the extension of existence and stability theorems to multivalued dynamical systems

$$\frac{\mathrm{d}z}{\mathrm{d}t} \in F(z)$$

where F is an upper hemicontinuous correspondence such that the image $F(z)$, of any point z in an open subset Ω of R^n, is a compact convex subset of R^n. For such systems, Lyapunov functions have been defined with the same purposes as for ordinary systems (see Champsaur et al. 1977; Aubin and Cellina 1984).

Lyapunov functions are used in many other economic models, to prove the convergence of non-tâtonnement processes for example (see Arrow and Hahn 1971) or to investigate the stability properties of a process of free entry and exit of firms, facing random demand and guided by expected profits (see Drèze and Sheshinski 1984). Of particular interest is the use of a Lyapunov function of the form $(Q - Q^e) \cdot (k - k^e)$, where k is the vector of capital stocks in the economy and Q is the vector of current prices for investment goods, to show that any optimal growth path tends to the (suitably modified) golden rule capital stock k^* when the discount rate is not too large (see Brock and Scheinkman 1976; Cass and Shell 1976).

Till now we have dealt only with dynamical stability, i.e. with qsts typically like the following one: two trajectories happen to pass through two neighbouring points; does it imply that they will ever after remain close to each other? Around 1970, G. Debreu and S. Smale introduced structural stability into economic theory, i.e. stability with respect to parameters of the system; a typical question is here: is the configuration of competitive equilibria of an economy (for example the fact that they are isolated) stable when the initial endowments of the agents in the economy change? Almost a century before, Poincaré introduced and systematically explored the concept of bifurcation in mathematics (see Poincaré 1881); the word came to him as a natural comparison with daily experience:

> On voit que les deux catégories d'ellipsoïdes forment deux séries continues de figures d'équilibre. Mais il y a une figure qui est commune aux deux séries et qui est, si l'on veut me permettre cette comparaison, un point de bifurcation. (Poincaré 1892, p. 810)

Bifurcation is a basic concept for the study of structural stability; even if the latter expression was to come much later, the essence of the approach is in Poincaré's works.

In the introduction (Poincaré 1882), Poincaré refers to 'les recherches ultérieures parmi lesquelles les plus importantes sont, sans contredit, celles de M. Liapounoff'. It seems indeed that no other mathematician of the time saw better than Lyapunov did the significance of Poincaré's new concepts and methods. In the three volumes (Lyapunov 1906–1912), Lyapunov explored in great detail the bifurcation of the equilibrium configurations of a rotating homogeneous mass of liquid. The ultimate goal was to explain the evolution of stars. As long as the angular velocity Ω of the rotating mass is less than or equal to a critical value ω^c, there is one and only one equilibrium configuration for each velocity ω, and it is an ellipsoid. But at ω^c a bifurcation appears: at ω^c the equilibrium configuration is still unique and is an ellipsoid but, in Lyapunov's own words, 'C'est l'ellipsoïde, par lequel on entre dans la série des figures d'équilibre que M. Poincaré a appelé pyriformes' (Lyapunov 1906–1912, vol. 3, p. 6). It is indeed shown (Lyapunov 1906–1912, vol. 1, pp. 216–17 and vol. 3, p. 106) that there exists an interval $[\omega^c, \overline{\omega}]$ such that, for every ω in this interval there are two equilibrium configurations: the usual ellipsoid and a pear-shaped configuration, whose symmetry and stability properties were systematically investigated by Lyapunov. This is a study in structural stability, the last one to appear before 1937, at which time Andronov and Pontrjagin (1937) picked up the subject, which has been exploding since then.

It has recently been shown that in (strictly deterministic) economic growth models, bifurcation phenomena can take place which are strikingly similar to those explored by Lyapunov: ω is replaced by the discount rate r, the ellipsoids by steady states and the pear-shaped configurations by closed cycles that bifurcate from the steady state for some value r_o of r (see Benhabib and Nishimura 1979). Bifurcations even appear in stationary competitive monetary economies: at critical values of some parameters of the economy – for example the degree of concavity of utility functions – a stationary equilibrium bifurcates towards a line (a 'série', in Poincaré's words) of stationary equilibria on one hand, and a simultaneous line of closed cycles; the latter are the business cycles of the model, and their stability under suitable assumptions has been shown using Poincaré–Lyapunov methods (see Grandmont 1985).

Economists tend to know Lyapunov for his celebrated functions, it appears that there is even more to interest them in the various approaches to stability that Lyapunov has developed during his lifelong study of dynamical systems.

See Also

- ▶ Correspondence Principle
- ▶ Gross Substitutes

Bibliography

Andronov, A.A., and L.S. Pontrjagin. 1937. Systèmes grossiers. *Doklady Akademii Nauk* 14: 247–251.

Arrow, K., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden- Day.

Aubin, J.P., and A. Cellina. 1984. *Differential inclusions*. Berlin: Springer.

Benhabib, J., and K. Nishimura. 1979. The Hopf bifurcation and the existence and stability of closed orbits in multisector models of optimal economic growth. *Journal of Economic Theory* 21: 421–444.

Brock, W.A., and J.A. Scheinkman. 1976. Global asymptotic stability of optimal control systems with applications to the theory of economic growth. *Journal of Economic Theory* 12: 164–190.

Cass, D., and K. Shell. 1976. The structure and stability of competitive dynamical systems. *Journal of Economic Theory* 12: 31–70.

Champsaur, P., J. Drèze, and C. Henry. 1977. Stability theorems with economic applications. *Econometrica* 45: 273–294.

Coddington, E.A., and N. Levinson. 1955. *Theory of ordinary differential equations*. New York: McGraw-Hill.

Drèze, J.H., and E. Sheshinski. 1984. On industry equilibrium under uncertainty. *Journal of Economic Theory* 33: 88–97.
Grandmont, J.M. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1045.
Hahn, W. 1967. *Stability of motion*. Berlin: Springer.
Lyapunov, A. 1892. Obshc˘aya zadac˘a ob ustoic˘ivosti dviz˘eniya (The general problem of the stability of motion). Kharkov Mathematical Society. The 1907 French translation has been reproduced in Annals of Mathematics Studies 17. Princeton: Princeton University Press, 1949.
Lyapunov, A. 1906–1912. *Sur les figures d'équilibre peu différentes des ellipsoïdes d'une masse liquide homogène douée d'un mouvement de rotation*. 3 vols. St. Petersburg: Académie impériale des Sciences.
Poincaré, H. 1881. Mémoire sur les courbes définies par une équation différentielle. *Journal de Mathématiques Pures et Appliquées* 7(3), 1881, 375–422; 8(3), 1882, 251–296; 1(4), 1885, 167–244; 2(4), 1886, 151–217.
Poincaré, H. 1882. Les formes d'équilibre d'une masse fluide en rotation. *Revue Générale des Sciences* 3: 809–815.
Rouche, N., and J. Mawhin. 1980. *Ordinary differential, equations: Stability and periodic solutions*, Surveys and reference works in mathematics. London: Pitman.
Rouche, N., Habets, P., Laloy, M. 1977. *Stability theory by Lyapunov's direct method*, Applied mathematical sciences, vol. 22. Berlin: Springer.

Lyapunov's Theorem

Peter A. Loeb and Salim Rashid

A result with numerous applications in economics is the theorem of Lyapunov (1940), which states that the range of a nonatomic totally finite vector-valued measure is both convex and compact. That is, let Σ be a σ-algebra in a set X and let μ_1, μ_2,. . .,μ_k be totally finite, nonatomic, signed measures on the measurable space (X, Σ). Then the range of the vector measure $\overline{\mu} = (\mu_1, \mu_2 \ldots \mu_k)$, that is, the set $\{\overline{\mu}(A) : A \in \sum\}$, is a convex, compact subset of R^k.

Proof: An elegant proof of Lyapunov's theorem due to Lindenstrauss (1966) but modified here for the case of signed measures, uses the measure $\mu = |\mu_1| + |\mu_2| + \cdots + |\mu_k|$ and the fact that the subset $W = \{g : 0 \leq g \leq 1\}$ of $L_\infty(\mu)$ is convex and also compact in the weak* topology, that is, the topology generated by $L^1(\mu)$. Here for $i = 1, 2, \ldots, k$, $\mu_i = \mu_i^+ - \mu_i^-$ is the Jordan decomposition of μ_i and $|\mu_i| = \mu_i^+ + \mu_i^-$; both μ_i^+ and μ_i^+ are absolutely continuous with respect to μ. Therefore, the mapping T from W into R^k obtained by setting

$$T(g) = \left(\int_x g\,d\mu_1^+ - \int_x g\,d\mu_1^-, \ldots, \int_x g\,d\mu_k^+ - \int_x g\,d\mu_1^-\right)$$

for each $g \varepsilon W$ is both affine and weak*-continuous. It follows that the range of T is a convex, compact subset of R^k. Given a vector υ in that range, the set $W_\upsilon = \{g \in W : T(g) = \upsilon\}$ is convex and also compact in the weak* topology. By the Krein–Milman theorem, W_υ contains extreme points. Lyapunov's theorem is established for R^k by showing that any extreme point in W_υ is the characteristic function χ_A of a set $A \varepsilon \Sigma$; intermediate result, in turn, is established with a proof by induction on the dimension k.

Lindenstrauss omits the proof for $k = 1$ because it is similar to the induction step. The assumptions for the induction step are that the result and therefore Lyapunov's theorem hold for dimension $k - 1$ and there is an appropriate vector υ in R^k and an extreme point g in W_υ such that g is not a characteristic function. Thus there is an $\varepsilon > 0$ and a set $Z \varepsilon \Sigma$ such that $\mu(Z) > 0$ and $\varepsilon \leqslant g \leqslant 1 - \varepsilon$ on Z. One may suppose that $\mu_1^+(Z) > 0$, other cases being similar. By reducing Z if necessary, one may further suppose that $\mu_1^-(Z) > 0$. Sice μ_1 is nonatomic, there is a measurable set $A \subset Z$ with $\mu_1(A) > 0$ and $\mu_1(Z - A) > 0$ if $\mu_1(A) = 0$ and $\mu_1(Z - A) = 0$ for $i = 2, 3, \ldots, k$ (or if $k = 1$), then B and C denote the empty set. Otherwise, by the induction hypothesis, there are measurable sets $B \subset A$ and $C \subset Z - A$ with $\mu_i(B) = (1/2)\mu_i(A)$ and $\mu_i(C) = (1/2)\mu_i(Z - A)$ for $i = 2, 3, \ldots, k$. There are numbers s and t, not both zero but each in the interval $[-\varepsilon, \varepsilon]$, such that if

$$h = s(\chi_A - 2\chi_B) + t(\chi_{Z-A} - 2\chi_C),$$

then $\int_x h\,d\mu_1 = 0$. Since $\int_x h\,d\mu_1 = 0$ for i = 2, 3, . . ., k, and $|h| \leqslant g \leqslant 1 - |h|$ on X, $g \pm h \in W_v$.

L

Since $h \neq 0$, this contradicts the assumption that g is an extreme point of W_v.

Applications

Extreme points appear not only in Lindenstrauss' proof of Lyapunov's theorem but also in the application of that theorem to the proof of the so-called 'bang-bang' principle; that is, objects can be satisfactorily controlled by using only extreme points. Models employing the bang-bang principle have been used in economics, but these models require assumptions that are usually too restrictive. An application of Lyapunov's theorem in statistics that is similar to the bang-bang principle (Dvoretzky et al. 1951) allows one to avoid randomization in hypothesis testing when the unknown distributions are at least known to be nonatomic. A slight modification of the same result shows the efficacy of a pure strategy in certain zero-sum twoperson games.

For most applications of Lyapunov's theorem in economics, it is enough to know that for a finite nonatomic measure $\overline{\mu}$ with values in the positive orthant of R^k, the closure of the range of $\overline{\mu}$ is convex. Extensions of Lyapunov's theorem can be found in the work of Robertson and Kingman (1968), who obtain a necessary and sufficient condition for Lyapunov's theorem to hold in the infinite dimensional case, and in the work of Armstrong and Prikry (1981), who establish an analogue of the convexity part of Lyapunov's theorem for finitely additive measures. Another extension, Loeb (1973), deals with nonstandard economies.

Using nonstandard analysis as developed by Robinson (Robinson 1966), one can construct natural models of large but finite economies in which the set of traders is an initial segment 1, 2, . . ., γ of the nonstandard natural numbers: here γ is an infinite integer. Each trader has an infinitesimal endowment in the goods of the economy. In this setting the following analogue of Lyapunov's theorem holds for a '*-finite' set of infinitesimal vectors $x_1, x_2, \ldots, x_\gamma$ in the nonstandard extension $^*R^k$ of R^k: let S denote the set of 'internal' subsets of the set of traders; then the set of sums

$$\left\{\sum_{i\in A} x_i : A \in S\right\}$$

is essentially convex; that is, it is convex except for infinitesimal errors.

Direct application of Lyapunov's Theorem to problems in economics was initiated in the seminal papers of Aumann (1964, 1966). Aumann's approach was extended by Schmeidler (1969), and a unified general treatment was provided by Hildenbrand (1974). The model for an economy employed by these authors is a nonatomic measure space (T, Σ, υ). The nonatomicity of the measure captures the idea that in a competitive market no single individual can unilaterally alter the market outcome. In a non-pathological market with a finite number of traders, some agent has a positive and therefore non-negligible influence. Only in a model with an infinite number of traders can all individuals be negligible. Given such a model, if $i(t)$ denotes the initial endowment of individual t, then a feasible allocation of commodities, with $x(t)$ denoting the assignment for trader t, has the property that

$$\int x(t)\upsilon(\mathrm{d}t) = \int i(t)\upsilon(\mathrm{d}t).$$

Lyapunov's theorem has been applied to economies modelled by nonatomic measure spaces to establish the equivalence between the set of core allocations and the set of competitive equilibria. That is, if x is an assignment of commodities and $P(x(t))$ denotes the set of points preferred by trader t to the point $x(t)$, then a set G which contains any point which at least one coalition prefers to its allocation $\{x(t)\}_{t\in S}$ is obtained by setting $G\int_{t\in T}(P(x(t)) \cup \{0\})\upsilon(\mathrm{d}t)$.

That is,

$$G\left\{\int_{t\in T} y(t)\upsilon(\mathrm{d}t) : y \in P[x(t)] \cup \{0\}\right\}.$$

If G is convex and the origin is at the boundary of G, then G is on one side of a hyperplane

through the origin. It is relatively straightforward to show that the normal to that hyperplane is a competitive equilibrium price vector. The convexity of the set G follows from Lyapunov's theorem. The equivalence theorem can be proved without the use of Lyapunov's Theorem, however, by showing that the convex hull of $\cup_{t \in T} P(x(t))$ is disjoint from the origin and then separating that convex hull from the origin to obtain the desired equilibrium price vector. This was the path followed by Aumann's original proof, which in turn followed the analogous proof for replicated economies by Debreu and Scarf (1963).

The Core Equivalence theorem has been refined with the following application of Lyapunov's theorem originating in the work of Schmeidler (1972). If the aggregate excess demand of a coalition sums to zero, then by Lyapunov's theorem there exist subcoalitions (subsets) of arbitrarily small measure for which the excess demand also sums to zero. Thus, if there exists one coalition which finds it feasible to break away and form an independent subeconomy, then there are arbitrarily small coalitions which also find it feasible to break away. Conversely, if an allocation is stable against the breaking away of large coalitions, it is also stable when the coalitions that can break away are arbitrarily small. The economic significance of this result stems from the fact that transactions and negotiating costs can become prohibitive as coalitions become very large; one need only, therefore, consider those (small) coalitions for which the negotiating costs are reasonable.

In proving the existence of a general equilibrium for nonatomic economies, Lyapunov's theorem is used to establish the conditions necessary for the applications of Kakutani's Fixed-Point theorem. Kakutani's result requires set-valued upper-semicontinuous mappings from a compact, convex set into itself. The sets which are the images under the mapping must be convex; Lyapunov's theorem is used to prove that one component of these sets, the aggregate excess demand, is convex. Similar proofs of the existence of equilibria existed before this use of Lyapunov's theorem, but they were founded on the somewhat awkward, and now unnecessary, assumption that individuals have convex preferences.

It is the desire for information about large but finite economies that motivates much of the research on the limiting infinite economies. In proving an appropriate convergence result for sequences of finite economies, one need not establish an exact equilibrium with aggregate excess demand equal to zero. It is sufficient to obtain an approximate equilibrium where the closure of the set of aggregate excess demands contains zero. For this reason, it is enough to know that for a finite nonatomic measure $\overline{\mu}$ with values in the positive orthant of R^k, the closure of the range of $\overline{\mu}$ is convex. Here also is the reason that the approximate convexity holding in nonstandard models is adequate for most results in economics.

See also

- Cores
- Large Economies
- Non-convexity
- Non-standard analysis
- Perfect Competition
- Shapley–Folkman Theorem

Bibliography

Armstrong, T.E., and K. Prikry. 1981. Liapounoff's theorem for non-atomic, bounded, finitely additive finite dimensional vector valued measures. *Transactions of the American Mathematical Society* 266: 499–514.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Aumann, R.J. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34: 1–17.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

Dvoretzky, A., A. Wald, and J. Wolfowitz. 1951. Relations among certain ranges of vector measures. *Pacific Journal of Mathematics* 1: 59–74.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Lindenstrauss, J. 1966. A short proof of Liapounoff's convexity theorem. *Journal of Mathematics and Mechanics* 15(6): 971–972.

Loeb, P.A. 1973. A combinatorial analog of Lyapunov's theorem for infinitesimally generated atomic vector measures. *Proceedings of the American Mathematical Society* 39: 585–586.

Lyapunov, A.A. 1940. On completely additive vector-functions. *Izvestiya Akademii Nauk SSSR* 4: 465–478.

Robertson, A.P., and J.F.C. Kingman. 1968. On a theorem of Lyapunov. *Journal of the London Mathematical Society* 43: 347–351.

Robinson, A. 1966. *Non-standard analysis*, Studies in Logic and the Foundations of Mathematics. Amsterdam: North-Holland.

Schmeidler, D. 1969. Competitive equilibria in markets with a continuum of traders and incomplete preferences. *Econometrica* 37: 578–585.

Schmeidler, D. 1972. A remark on the core of an atomless economy. *Econometrica* 40: 579–580.